Mutual Information analysis for feature and sensor subset selection in surface electromyography based speech recognition

Vivek Kumar Rangarajan Sridhar, Rohit Prasad, Prem Natarajan
Raytheon BBN Technologies
10 Moulton Street
Cambridge, MA 02478

vsridhar@bbn.com, rprasad@bbn.com, pnataraj@bbn.com

Abstract
In this paper, we investigate the use of surface electromyographic (sEMG) signals collected from articulatory muscles on the face and neck for performing automatic speech recognition. While previous work has typically used full-scale recognition experiments to evaluate appropriate feature representation schemes for sEMG signals, we present a systematic information-theoretic analysis for feature selection and optimal sensor subset selection. Our results indicate that Mel-cepstral frequency features are best suited for sEMG-based discrimination. Further, the sensor subset ranking obtained through the mutual information experiments are consistent with the results obtained from hidden Markov model based recognition. The framework presented here can be used for determining the best feature and sensor subset for a given speaker a posteriori, instead of determining them a priori from recognition experiments. We achieve a mean recognition accuracy of 80.6% with the best 5 sensor subset chosen by the MI analysis in comparison with 79.6% obtained from using the complete set of 11 sensors.

Index Terms: Surface electromyography, mutual information, hidden Markov models

1. Introduction
State-of-the-art automatic speech recognition technology can achieve impressive recognition performance on clean acoustic speech. While the performance typically drops due to speaker, channel and environmental variability, a variety of adaptation and discriminative transform techniques have been proposed to ameliorate these effects. However, these systems typically use parameterizations of the input acoustic speech signal without explicitly considering the underlying speech production process that generates the speech. Capturing and modeling the articulatory process that generates the acoustic signal can potentially model the true co-articulatory effects approximated using context-dependent models in traditional hidden Markov model (HMM) based speech recognition [1] and provide robustness in the presence of environmental noise [2].

Several methodologies for using speech production information in speech recognition have been investigated in the past. These include magnetic resonance imaging (MRI), electromagnetic midsaggital articulography (EMA) [3], discrete knowledge-based representations that describe articulation [4] or mechanics, etc. In this paper we address the problem of using surface electromyographic (sEMG) signals for speech recognition. Surface electromyography is the process of recording the electrical activity of a muscle. Since the myoelectric signal generated through muscle contractions is immune to acoustic noise, it can either augment or serve as an alternative to typical acoustic signals used in speech recognition. They also have the potential to facilitate secure form of communication as the myoelectric signals can still be recognized when the speaker speaks inaudibly.

Speech recognition using surface electromyographic signals has been addressed in past work for a variety of small vocabulary tasks. These tasks range from digit recognition [5, 6, 7], limited vocabulary word recognition of the order of 15-65 words [2, 8] and continuous speech recognition [9]. Both HMM-based recognition [5, 6, 8] as well as discriminative modeling techniques such as neural networks [2] have been used in previous work. These systems have demonstrated impressive recognition (classification) accuracies ranging from 68%-93% for limited vocabulary and digit recognition tasks.

In this work, we are interested in addressing the following outstanding issues in sEMG-based speech recognition:

1. Features such as Mel-frequency cepstral coefficients (MFCCs), time-domain features, wavelet features have previously been used for sEMG-based recognition. While previous work has evaluated the usefulness of these features by performing full-scale recognition experiments, performing a systematic analysis of the informativeness of each feature can provide meaningful insights into suitability of these features for recognition.

2. There has been lack of a systematic analysis of which sensors provide maximum discriminative information. Previous work has addressed sensor subset selection by performing recognition experiments on the combinations of all sensors, a computationally expensive methodology [10]. Further, determining the optimal sensor subset for a speaker a priori can aid in maximizing the recognition performance with less computational overhead.

We perform the feature and sensor subset analysis using mutual information (MI) as the fidelity metric. First, we present a framework using MI to quantify the information contained in the feature stream to discriminate between the classes (words). Our objective is to use MI to evaluate the usefulness of several feature representation schemes. Second, we use the MI analysis to systematically select the optimal subset of sensors for sEMG-based speech recognition. Finally, we ratify the results of our mutual information analysis using HMM based recognition experiments.

2. Data
The sEMG data used in this work were collected from 9 subjects (5 males and 4 females). All the subjects were native American English speakers. Eleven sEMG sensors were positioned across...
6 anatomical regions (supralabial, labial, sublabial, submental neck, middle neck, and lateral neck) to capture optimal speech-related muscle movement. Detailed information regarding the apparatus and methodology used in the data collection can be found in [8].

The sEMG data used in the experiments were collected under two conditions, voiced and mouthed. In the vocalized mode, the subjects speak the word audibly whereas in the mouthed mode only the muscular action involved in speaking is performed. Each subject was made to read out 65 words, three times. The data collection was performed in two different sessions, with three instances of each word spoken continuously in each session. We used user generated triggers as the markers for segmenting the continuously spoken words. The training set comprised of four instances of each word (instances 1 and 3 from the two recorded sessions) and the test set consisted of 2 instances (instance 2 from each session).

3. Mutual Information analysis of sEMG signals

In this section, we present a framework with mutual information as a fidelity metric, to measure the amount of information in a given feature vector for discriminating between the target classes (words or phoneems). Let $c$ denote the word class and $X$ denote the $D$-dimensional feature vector extracted from the sEMG signal, i.e., $X \in \mathbb{R}^D$. The MI between $X$ and $c$ is defined by

$$I(X; c) = \sum_c P_c \int_X p(X|c) \log \left( \frac{p(X|c)}{p(X)} \right) dX$$

(1)

where $P_c$ denotes the prior probability of class $c$.

In reality since neither $p(X)$ nor $p(X|c)$ are known exactly, we typically assume some form of parametric distribution (Gaussian) for the densities and estimate the parameters from the data. While assuming a Gaussian distribution is a good approximation to the true distribution, it is not possible to develop a closed form solution for MI [11] or numerically evaluate the MI. In this work, we assume the class conditional distribution as a discrete distribution, i.e., we quantize the feature observation space with a finite number of quantization bins. This is equivalent to partitioning the feature space into nonoverlapping regions, and approximating the joint probability density function, $p(X, c)$ within each region, by a uniform function. Let us denote $Q(.): \mathbb{R}^D \to A_c$, $|A_c| < \infty$, as the quantization function, the MI can be approximated as

$$I(Q(X); c) = \sum_c P_c \sum_{q \in A_c} p(Q(X) = q|c) \log \frac{p(Q(X) = q|c)}{p(Q(X) = q)}$$

(2)

where $p((Q(X) = q)\mid c)$ is a discrete nonparametric distribution. A similar quantity can be computed for any new feature vector, $Y$. Then MI across the two features can be compared if we use the same number of codewords $|A_c|$ in the quantization process. The approximation of MI using quantization provides a computationally inexpensive diagnostic tool that can be used to evaluate different feature vectors for sEMG-based recognition before building a recognition system in the new feature space.

For computing the quantized mutual information (QMI) for different feature representations, we collected feature vectors for each class from the entire data and used K-means clustering to cluster the feature vectors in each class into 10 diagonal Gaussians, resulting in a total of 650 codewords ($10 \times 65$ classes) across all classes. A likelihood metric was then used to label each feature vector with the label of the closest Gaussian. Subsequently, Eq. (2) was used to compute the MI between the class and quantized feature vectors. We also tried a K-means approach that used the centroid of each cluster as codeword and a distance-based metric for computing the closest labels. In this work, we present results using the diagonal Gaussians only.

4. MI analysis for feature selection

We analyzed the suitability of MFCC, time-domain (TD), short-time Fourier transform (STFT) and wavelet transform (WT) features for sEMG-based recognition. We used the entire training data in the analysis.

The sEMG data was first downsampled from 20 KHz to 1 KHz, and the DC-offset was removed from each channel. We extracted 6 MFCCs at 25 ms frame rate with 50ms overlapping window for each of the 11 sEMG channels. Time domain features were extracted using the original 20 KHz signal as previous studies have demonstrated the utility of high-frequency TD features [9]. The TD features included the mean absolute value, mean absolute value slope, zero crossing, slope sign change and waveform length [12]. STFT analysis on the 1 KHz downsampled signal with 50 ms observation window and 25 ms frame shift resulted in 13 coefficients per frame for each channel. For extracting wavelet features, we used wavelet packets iterating a

<table>
<thead>
<tr>
<th>Subject</th>
<th>Audio</th>
<th>TD</th>
<th>STFT</th>
<th>WT</th>
<th>MFCC</th>
<th>Audio</th>
<th>TD</th>
<th>STFT</th>
<th>WT</th>
<th>MFCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject 1</td>
<td>4.324</td>
<td>0.389</td>
<td>1.233</td>
<td>1.582</td>
<td>2.060</td>
<td>0.618</td>
<td>0.750</td>
<td>0.221</td>
<td>1.694</td>
<td></td>
</tr>
<tr>
<td>Subject 2</td>
<td>5.037</td>
<td>0.907</td>
<td>1.437</td>
<td>0.300</td>
<td>2.271</td>
<td>1.505</td>
<td>1.800</td>
<td>2.155</td>
<td>2.671</td>
<td></td>
</tr>
<tr>
<td>Subject 3</td>
<td>4.243</td>
<td>0.448</td>
<td>0.484</td>
<td>0.751</td>
<td>1.528</td>
<td>0.641</td>
<td>0.317</td>
<td>0.468</td>
<td>1.791</td>
<td></td>
</tr>
<tr>
<td>Subject 4</td>
<td>2.397</td>
<td>0.786</td>
<td>0.915</td>
<td>1.514</td>
<td>1.870</td>
<td>1.004</td>
<td>0.896</td>
<td>0.322</td>
<td>1.955</td>
<td></td>
</tr>
<tr>
<td>Subject 5</td>
<td>4.849</td>
<td>0.669</td>
<td>1.230</td>
<td>1.837</td>
<td>2.005</td>
<td>0.814</td>
<td>1.221</td>
<td>1.616</td>
<td>2.183</td>
<td></td>
</tr>
<tr>
<td>Subject 6</td>
<td>4.430</td>
<td>0.795</td>
<td>0.638</td>
<td>0.123</td>
<td>1.449</td>
<td>0.622</td>
<td>0.602</td>
<td>0.231</td>
<td>1.439</td>
<td></td>
</tr>
<tr>
<td>Subject 7</td>
<td>3.783</td>
<td>0.214</td>
<td>1.342</td>
<td>0.780</td>
<td>1.462</td>
<td>0.438</td>
<td>1.160</td>
<td>1.289</td>
<td>1.531</td>
<td></td>
</tr>
<tr>
<td>Subject 8</td>
<td>4.059</td>
<td>0.673</td>
<td>1.393</td>
<td>1.406</td>
<td>1.753</td>
<td>0.561</td>
<td>1.420</td>
<td>1.040</td>
<td>1.862</td>
<td></td>
</tr>
<tr>
<td>Subject 9</td>
<td>4.222</td>
<td>0.710</td>
<td>0.435</td>
<td>0.593</td>
<td>1.447</td>
<td>0.638</td>
<td>0.330</td>
<td>0.161</td>
<td>1.620</td>
<td></td>
</tr>
</tbody>
</table>
Daubechies two channel filter bank block (db4) with balanced full tree-structure. The average coefficient at each node of the filterbank was used as the feature vector. For a tree depth of 3, this resulted in 8 coefficients per frame for each channel. Finally, since the voiced portion of the data is accompanied with audio recordings, we extracted standard MFCC features used in traditional automatic speech recognition (14 MFCC features were extracted from overlapping frames of audio data at the rate of 100 frames per second) from the audio data.

The QMI between the classes (words) and different feature representations is shown in Table 1. From Table 1, we can infer that using MFCC features from the audio signal (traditional ASR) provides maximum discrimination among the classes. For sEMG-based features, MFCCs contain the maximum amount of discriminative information between the classes followed by WT, STFT and TD features. Even though MFCCs are not the most intuitive features for myoelectric signals (the nonlinear warping used in MFCC is associated with the acoustic perception of loudness), they offer good discrimination between the classes for sEMG based recognition. These results conform to the empirical recognition experiments presented in [13]. To our knowledge, the MI analysis reported here is the first systematic analysis of various feature representations for sEMG signals on a single corpus. Previous work has used different features for a variety of data sets and hence has not considered the usefulness of features across one data set. The framework presented here can also be used a tool for designing features that have maximal discrimination across sessions and speakers, the ultimate goal of sEMG-based recognition.

5. MI analysis for sensor subset selection

sEMG signal data are typically collected by placing sensors on the major articulator muscles. While there is no consensus regarding the placement of the sensors for optimal signal acquisition with the end goal of sEMG-based recognition, five to six sensors mounted on parts of the face and neck have been used in previous work [7, 6, 13]. In the data used in our experiments, 11 sensors were used for collecting the sEMG signals. Using the data from all the 11 sensors in recognition experiments can lead to data sparsity due to very high dimensional feature vectors generated by combining the features from all the sensors. Further, different speakers may use a subset of the sensors during speech production. Sensor subset selection in [10] was performed in a computationally expensive framework, where all the combinations of sensors were used in recognition experiments and the best subsets were chosen by evaluating the a posteriori recognition performance. In order to understand and investigate the nature of speech production in different individuals, we carried out the MI analysis on subsets comprising 5 sensors. However, in order to reduce the combinatorial number of experiments (11 choose 5), we fixed sensor 1 and 2 constant across all the subsets as they demonstrated the best performance in pilot experiments. This resulted in 84 subsets. We used MFCC features along with the deltas as the features, based on our feature selection experiments. Figure 1 shows the placement of the sensors on a subject.

Table 2 shows the top and bottom sensor subsets in terms of mutual information for the mouthed condition. It is interesting to note that the top sensor subset always has a combination of sensors from the neck and face (two sides). On the contrary, the bottom sensor subsets comprise redundant sensors from the neck. Comparison with QMI experiments for the 11 channel condition with the end goal of sEMG-based recognition, five to six sensors from the neck and face (two sides). On the contrary, one can use a similar subset of sensors on a subject to achieve optimal performance.

6. HMM recognition experiments

In this section, we present HMM-based recognition experiments based on our initial analysis. We used the BBN Byblos hidden Markov model system [1] for performing the recognition experiments. Based on the results of preliminary feature analysis, we extracted 6 MFCCs at 25 ms frame rate with 50ms overlapping window for each of the 11 sEMG channels. Along with the delta coefficients, this resulted in a 132 dimensional feature vector. Each word in the vocabulary was represented using a 10-state HMM. The HMM topology consisted of 10 states and the transition matrix was defined such that the model either stayed in the same state, or advanced either one or two states between consecutive frames. The output distribution is a mixture of diagonal Gaussians. Estimation of the HMMs occurs in multiple steps, involving the K-means algorithm to generate an initial set of Gaussians and the EM algorithm to generate the mixture weights and re-estimate the Gaussians. Based on the results of the sensor subset experiments, we repeated the HMM-based recognition experiment for the top and bottom sensor subsets. We used 1 Gaussian per word per dimension of the feature vector, i.e., 128 Gaussians for the experiments using all the sensors and 64 Gaussians for the experiments using the 5 sensor subset.

The results in Table 2 show that the trend observed in the MI experiments holds for the HMM experiments. The recognition accuracies from the top sensor subset is quite close to that achieved with all the 11 sensors. Essentially, the MI analysis for choosing the best sensor subset eliminated the most redundant sensors. The mean recognition accuracy across the 9 subjects is 79.7% while using all the 11 sensors, and 80.6% with

\[ \Delta I(Y; X) = I(Y; c) - I(X; c) = I(x; c|X) \]

As \( I(x; c|X) \) is constrained to be positive, adding additional features to the baseline, must not reduce the MI
the top sensor subset for each speaker\(^1\). Sensor subset experiments using MI across all the subjects resulted in \(\{1,2,6,8,9\}\) as the top sensor subset and achieved a mean recognition accuracy of 81.8%. Hence, one can a priori fix the optimal set of sensors on a given subject without having to perform recognition experiments and then determine the best sensor subset \textit{a posteriori} [10]. Our framework is a robust method for both feature and sensor subset selection. It can also provide insights into the speaker specific use of articulatory muscles involved during speech. For example, one can determine that a speaker predominantly uses his/her facial muscles more than neck muscles while speaking. Our scheme could also potentially find use in literacy assessment and guidance.

7. Conclusion and Future Work

We presented a mutual information framework for evaluating the informativeness of several feature representations of sEMG signals. The results of our analysis demonstrate that MFCC features are the most useful in discriminating among the word classes. Our framework can be used as a diagnostic tool for feature selection, obviating the need to perform computationally expensive full-scale recognition experiment. We extended the MI framework to select the best sensor subset. The MI experiments were ratified by performing HMM-based recognition experiments. Our experiments indicate that speakers use different set of articulatory muscles and exploiting this information can be beneficial in maximizing sEMG-based speech recognition accuracy. As part of future work, we are exploring normalization and adaptation techniques for performing session independent and speaker independent sEMG-based recognition.

8. Acknowledgments

This study was funded by the United States Defense Advanced Research Projects Agency (DARPA) under contract W15P7T-05-P218. The views, opinions, and/or findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

\(^1\)The results are statistically insignificant at \(\alpha = 0.01\) as determined by a difference of proportions test.

9. References


