Towards a practical silent speech recognition system

Towards a Practical Silent Speech Recognition System

Yunbin Deng [1], *James T. Heaton,*[2] *Geoffrey S. Meltzner* [1]

[1] BAE Systems, Burlington, MA, USA
[2] Department of Surgery, Massachusetts General Hospital, Boston, MA, USA
{yunbin.deng, geoffrey.meltzner}@baesystems.com, james.heaton@mgh.harvard.edu,

Abstract

Our recent efforts towards developing a practical surface electromyography (sEMG) based silent speech recognition interface have resulted in significant advances in the hardware, software and algorithmic components of the system. In this paper, we report our algorithmic progress, specifically: sEMG feature extraction parameter optimization, advances in sEMG acoustic modeling, and sEMG sensor set reduction. The key findings are: 1) the gold-standard parameters for acoustic speech feature extraction are far from optimum for sEMG parameterization, 2) advances in state-of-the-art speech modelling can be leveraged to significantly enhance the continuous sEMG silent speech recognition accuracy, and 3) the number of sEMG sensors can be reduced by half with little impact on the final recognition accuracy, and the optimum sensor subset can be selected efficiently based on basic mono-phone HMM modeling.

Index Terms: silent speech interface, sEMG signal processing and feature extraction, sEMG speech recognition.

1. Introduction

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

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

Several recent research efforts have focused on developing this technology. Chan et al. [3] obtained a 93% recognition rate on a vocabulary of 10 digits using 5 sEMG channels on the face and neck for vocalized speech. Jou et al. [5] further extended the vocabulary size to 108 words but at the cost of reduced recognition accuracy (68%). Lee [6] was able to achieve a mean 87% recognition rate on 60 vocalized words for 8 male, Korean speakers. Wand and Schultz have pushed sEMG-based recognition towards continuous speech recognition [7], ultimately achieving a Word Error Rate (WER) of 15.66% on a vocabulary of 108 words [8]. More recently, Meltzner et al. reported 96.9% recognition rate on a continuous vocabulary of 200 words [10].

Although non-acoustic ASR performance has significantly improved, it still vastly lags that of standard ASR, which prevents it from becoming a viable and practical technology. We have explored numerous signal processing approaches to improve sEMG-based silent speech recognition accuracy. Specifically, we have focused applying speaker dependent optimal sEMG feature parameterization as well as adapting the most recent major advances in acoustic speech modelling towards the silent speech recognition domain. We report the results of these efforts below.

2. Data Collection and Experimental Setup

2.1. Corpus Design

State of the art sEMG processing techniques can only achieve high recognition accuracy on relatively small vocabularies, yet an unlimited, or at least much larger vocabulary is needed for flexible deployments. To enable this capability, we designed a silent speech data corpus that covers commonly used English words (about 2000) and balances the frequency of phoneme combinations for unforeseen testing words. Our designed corpus consists of five subsets: 1) **TIMIT data**: 150 utterance from the SI portion and 450 utterances from the SX part were chosen [11]. This portion of data was selected to balance the frequency of phoneme combinations. 2) The **Special Operations** data set, which was designed for use in military scenarios includes the 26 letters of the NATO alphabet and a set of text commands equivalent to hand signal commands [12]. 3) The **Common Phrases** set contains 100 most commonly used English phrases [13]. 4) The **Text Message** set is of frequently used text messages extracted from [14], which would be used to demonstrate hand-free silent text messaging for commercial applications. And finally, 5) the **Digits and Date String** set is a hand crafted data set to cover phone numbers, names, and dates.

The overall data collection corpus contained a total 1200 utterances that was collected in a two-hour session for each subject. We then conducted testing on each of the four subsets with models trained on the other three subsets and the common TIMIT subset. Half of the data from the common phrase and text messages sets were reserved for cross validation. The validation set contained 199 sentences and can be used for frontend parameter optimization and future work on task adaptation and neural network modeling cross validation. Table 1 lists the vocabulary size of each test set and the number of word not seen in its training set.

Table 1. *Text statistics of the four subtasks: S-Special Operation, C-Common phrases T-Text Messages, D-Digits and Date String*

Tasks	S	C	T	D
Vocabulary	154	167	115	88
Unseen Words	70	36	36	41
# of sentences	138	246	152	90

2.2. Data Collection

For the above developed text corpus, six adult native speakers of English were recruited (two male, four female). The 1200 utterances were collected from each subject. The data collection protocol for this continuous mouthing mode was similar to that of the isolated word data collection described in [2]. A set of 8 wireless sEMG sensors were used with the sensors being placed a strategic locations on the face and neck. The sensors are a customized version of the Trigno™ wireless sensors (Delsys Inc., Natick MA) and are attached to the skin with double-stick adhesive strips. The sEMG signals were sampled at 20kHz per channel. More details about the data collection protocol can be found at [15].

3. sEMG Feature Parameterization

Traditional acoustic speech recognition systems use MFCC features for signal parameterization. Our previous studies have shown that modifying the MFCC parameterization algorithm to account for the characteristics of sEMG signals results in performance that exceeds that of using the standard MFCC algorithm, as well as that of other feature sets, including time domain features, spectral features and wavelet features [16][17][18]. In addition, previously reported sEMG speech recognition systems have typically used window sizes of around 30 ms with a window shift of approximately 10 ms, a typical scheme for traditional speech feature extraction [16]. Conversely, other studies have reported using a larger 54ms window with a much shorter 4ms overlap [19]. The large variability in windowing parameters indicates that the best approach has yet to be determined.

3.1. Impact of Window Size and Shift

For this line of investigation, we used the data from 2 of the 6 subjects. We concurrently varied the window size between 25-40ms and the window shift size between 10-20ms in 5ms increments. A basic mono-phone model was used for the recognition back end. Table 2 summarizes the impact of window size and shift parameters on the speech recognition accuracy for each subject. It is clear that recognition accuracy is sensitive to these parameters, and the typical parameters used for acoustic speech (30ms/10ms, shown in red color) are far from optimum for sEMG-based speech recognition. This could be due to the fact that speakers have highly variable speaking rates (after silence removal, the recorded speech varies from 1.1 hour to 1.7 hour for the same 1200 utterances) and signal statistics under the silent speaking mode.

Table 2. *The window size and shift impact on subject 1and subject 2 speech recognition WER (%).*

Subject	Size\shift (ms)	10	15	20
1	25	16.3	9.6	8.4
	30	**16.7**	8.7	8.5
	35	18.1	9.1	***7.0***
	40	13.2	8.9	8.0
2	25	40.9	31.2	44.2
	30	**39.0**	30.6	41.7
	35	37.0	29.8	40.3
	40	38.3	***28.2***	38.4

A longer window shift could result in loss of cepstral feature information, but it helps recognition accuracy in some cases. This is true when window shift increases from 10 ms to 15 ms, although not necessary when the shift increases from 15ms to 20 ms. These results may appear counter intuitive, but they could be attributable to the window shift effect on the delta cepstral feature, which are computed based on 5 sequential frames. A longer window shift could help capture the sEMG feature dynamics at a longer time frame for delta features. Thus, the window parameters make a tradeoff between the sEMG cepstral feature and delta feature information contribution to the final recognition accuracy. These results suggest that the sEMG signals need a long window and/or window shift compared with the acoustic speech signals. This is possibly due to the fact that acoustic speech has a richer spectrum content, and thus depends less on the context frames as compared with the sEMG counterpart.

3.2. Optimal Speaker Window Parameters

The results in Table 2 suggest that the optimal window parameters may be speaker dependent. As such we developed a means to adaptively determine the optimal window parameters in the 12 given combinations for each subject. To validate this procedure, we used the cross validation data (mono-phone model) to find the best performing window parameters for each speaker. Table 3 compares the performance of two uniform window parameter settings with the speaker optimized parameter setting and shows that the performance gain produced by speaker-adaptive window parameters is substantial.

Table 3. *Impact of FFT window parameters on WER (%). B-Baseline with 30ms window and 10ms shift; U-Uniform with 30ms window and 15ms shift; O-Optimum speaker-adaptive window parameter*

Subject\Settings	B	U	O
1	16.7	8.7	7.2
2	39.0	30.6	23.4
3	38.5	31.5	21.2
4	38.4	30.1	17.1
5	17.7	13.8	8.0
6	64.8	54.5	43.2
Mean	35.8	28.2	20.0

4. Advances in sEMG Speech Recognition

In this section, we detail our silent speech recognition experimental design, advances in sEMG acoustic modelling, and sEMG sensor set reduction.

4.1. sEMG Speech Detection and Feature Extraction

The sEMG signals were first subject to DC offset removal and down sampled to 5 kHz. Although there is little work reported in sEMG-based speech activity detection, we have developed a real-time robust silent speech activity detection algorithm based on a two-stage finite states machine utilizing multiple sEMG channels [10]. For feature extraction, we used a Hamming window with window size and shift adapted for each speaker, followed by cepstral analysis to have 7-dimension cepstral feature. 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.

4.2. Adapting Recent Advances in ASR

Although our parameterization optimization resulted in notable improvements in silent speech recognition performance, the resulting accuracy was still not satisfactory for practical applications. To further increase performance, we sought to adapt recent advances in the field of acoustic speech modeling to the silent speech recognition domain. In this section, we first describe our baseline systems and then detail the development of more advanced components, including data driven triphone modeling, heteroscedastic linear discriminant analysis (HLDA), maximum likelihood linear transformations (MLLT), and subspace Gaussian modelling (SGMM).

4.2.1. *The HTK vs. KALDI Baseline Systems*

The HTK toolkit [20], which has been the basis for our previously reported work [2][10], in not capable of supporting the implementation of more advanced recognition algorithms. As such, we switched to the KALDI speech recognition toolkit [21], which has been under active development for implementing many advanced algorithms.

As a first step, to verify that the switching toolkits would not degrade recognition performance, we compared these two systems with the very basic mono-phone modelling approach on two subjects. Both systems used 40 context-independent phonemes, with standard left-to-right three-state architecture. The HTK system models each state by 16 mixtures of Gaussian, or total 1920 Gaussians. The KALDI system contains a similar total of 1800 Gaussians, but each state is allowed to have a variable number of Gaussians determined by the training process. Table 4 summarizes baseline system performance. These results suggest that the KALDI baseline performance exceeds that of HTK, and, more importantly, that having variable number of Gaussian for each state vastly improves performance and is critical for sEMG-based speech recognition. Having established that the KALDI toolkit is sufficient for silent speech recognition, we proceeded to exploit the advanced algorithms contained within KALDI. All studies in the following sections are based on the KALDI toolkit.

Table 4. *The HTK vs KALDI baseline mono-phone system recognition WER (%). S-Special operation, C-Common phrases, T-Text messages, D-Digits and date strings*

Systems	Subjects	Tasks				Mean
		S	C	T	D	
HTK	1	24.9	26.3	29.6	19.5	25.4
	2	39.6	43.3	29.0	65.4	44.3
KALDI	1	10.2	8.1	4.2	11.9	8.7
	2	21.0	29.5	30.4	41.5	30.6

4.2.2. *Context Dependent sEMG Models*

State-of-the-art sEMG-based speech recognition systems work well only for small vocabulary tasks. To deploy a trained system for a new task, recent work has started to explore sub-word sEMG modelling [15][19]. Previous work explored phoneme and syllable-based modelling and reported WER of 37.6% for unseen words on a 32 word task [19]. A context dependent bundled phonetic feature model was also proposed for sEMG modeling and achieved average WER of 31.49% on a 101-word task [22].

As sEMG training data are very limited, we take the well-known data driven decision tree clustering approach to generate linguistic questions and synthesize the unforeseen triphones [21]. Our final triphone system has 500 tied states.

4.2.3. *Feature Adaptation and Dimension Reduction*

The high-dimensional multi-channel sEMG feature set are highly correlated and redundant. We thus applied the well-known HLDA feature dimension reduction technique and MLLT feature adaptation to enhance the discriminative power of the feature set [23][24]. HLDA utilizes 4 left frames and 4 right frames as context. The 112-dimension input feature was transformed into a 30-dimension discriminative feature space. Table 6 summarizes the recognition performance of the triphone recognition system with HLDA and MLLT. Compared with the mono-phone baseline (see Table 3 last column), this is an average relative WER reduction of 31.5%.

Table 5. *The triphone system with HLDA and MLLT recognition WER on silent speech (%)*

Subject\Task	S	C	T	D	Mean
1	2.3	1.4	4.2	5.0	3.2
2	7.1	15.6	21.0	24.3	17.0
3	16.7	12.9	13.1	30.8	18.4
4	11.3	2.8	10.3	21.0	11.3
5	3.0	2.4	2.8	15.7	6.0
6	16.5	20.9	37.4	29.8	26.1
Mean	9.5	9.3	14.8	21.1	**13.7**

4.2.4. *Subspace Gaussians Mixture Model (SGMM)*

Under the SGMM approach, all phonetic states share a common GMM structure, and the means and mixture weights vary within a subspace [25]. This allows a more compact representation and often gives better results with smaller amounts of training data, as is the case for silent speech recognition. In this experiment, the common GMM was trained with 200 mixtures. The final SGMM model had 800 leaves and 1200 sub-states. The SGMM model achieved relative WER reduction of 18.5% compared to the triphone baseline (see Table 6).

Table 6. *The SGMM (on top of Triphone+HLDA+ MLLR) system WER on silent speech (%)*

Subject\Task	S	C	T	D	Mean
1	2.0	0.0	0.9	2.7	1.4
2	8.0	15.4	15.9	15.4	13.9
3	13.6	5.2	12.1	20.7	12.9
4	10.6	3.7	10.3	18.2	10.7
5	3.4	1.2	5.6	12.2	5.6
6	12.7	15.8	35.1	26.2	22.4
Mean	8.4	6.9	13.3	15.9	**11.2**

4.3. sEMG Sensor Set Reduction

For a practical silent sEMG-based speech interface, the sEMG sensors need to be easily and reliably deployed, introduce minimal interference with recording location mobility, and remain relatively unobtrusive. Towards this goal, we have tested multiple hardware iterations with smaller and smaller sensor sizes, and in parallel have worked on reducing the number of sensor locations without causing substantial reduction in recognition accuracy. In this paper we describe results relating to our sensor number minimization efforts.

Our initial sEMG-based ASR system had 11 sensors, including 4 on the face and 7 on the neck/chin surface. We subsequently determined that this number could be safely reduced to 8 (4 on the face and 4 on the neck/chin surface) based on a sensor subset selection study, with less than 0.5% loss in recognition accuracy [26]. However, our current 8-sensor configuration can likely be reduced even further (with only moderate loss of recognition accuracy), based on the earlier sensor selection results. Our goal in the present study was to identify several 4-sensor subsets producing the least drop in recognition accuracy.

An exhaustive search of optimal sensor subsets based on complex acoustic modelling and system performance validation would be computationally expensive (i.e. it would require running 70 possible sensor combinations experiments). Instead, we chose an efficient and effective method to identify the optimum sensor subset based on the very basic mono-phone model, and then validating the results on a few top combinations using the most advanced acoustic modelling approaches.

Figure 1 illustrates the current eight-sensor configuration and the top six performing four-sensor subsets based on mono-phone model evaluation. Table 7 cross-validates the performance of the top performing sensor subsets using the advanced SGMM modeling approach. It shows that the performance of sensor sets 1, 2, and 5 are in the same order no matter whether we use the very basic mono-phone model or the most advanced SGMM model, thus the basic model can be used for much more efficient sensor subset selection. In addition, the best four-sensor subset achieved WER of 15.2%, which is only a 4% absolute drop from the full eight-sensor configuration. The top six subsets all tend to cover all four facial and neck areas (1. above the oral commissure, 2. below the oral commissure, 3. on the submental surface, and 4. on the ventral neck surface). These results make intuitive sense, because such sensor location combination captures more articulation information than otherwise.

Table 7. *Cross validation of sensor subset performance, mono-phone vs. SGMM, WER (%)*

System\Sensor Set	1	2	5
Mono-phone	31.1	32.8	34.8
SGMM	15.2	16.5	17.2

5. Discussion & Conclusions

Advances in sEMG signal processing, feature extraction, and sub-word modelling has continued to push the state-of-the-art in sEMG silent speech recognition technologies towards a practical, useful level. This study shows that flexible vocabulary silent speech recognition is achievable with phoneme models and by leveraging progress made in acoustic speech recognition. The current technology results in an average WER at the 10% level, while the best performing speaker achieved a WER as low as 1.4%. In addition, the sEMG sensor can be reduced to 4 with little drop in recognition accuracy.

The key limitation of current sEMG-based speech recognition technology is that high performance is achieved only with speaker dependent models. This is partially due to the large variation of the sEMG signals among subjects, and partially due to the very limited data available for speaker-independent sEMG model training. This problem could potentially be mitigated by applying the current speaker-dependent technology to a larger pool of subjects to accumulate the training data for speaker independent systems. Future work includes designing a flexible sEMG sensor patch for convenient deployment, and implementing the data collection, signal processing and recognition system on a small, portable device.

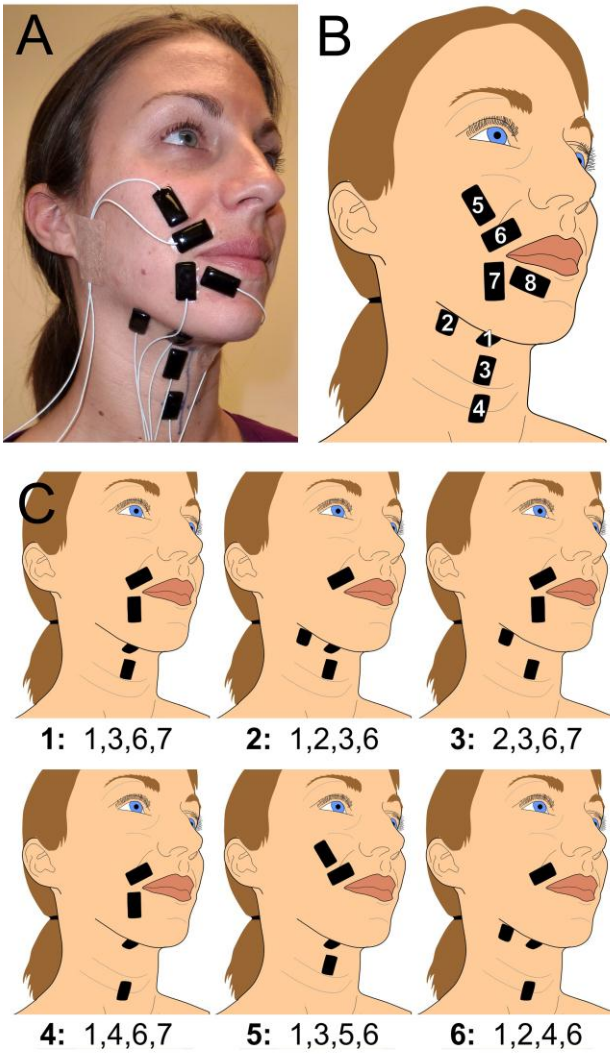

Figure 1. A) Photo of the complete sEMG sensor set, B) Drawing of the sensors with number labels, and C) graphic and numeric indication of the top six performing four-sensor subsets based on the mono-phone model.

6. Acknowledgements

The authors would like to thank Carlo J. De Luca, Gianluca De Luca, Don Gilmore, and Serge Roy of Delsys, Inc. for developing and providing the sEMG sensors and associated hardware. This study was sponsored by the United States Defense Advanced Research Projects Agency (DARPA) Defense Science Office Program. The views and conclusions in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of DARPA or the U.S. Government.

7. References

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

[2] Meltzner, G.S., Sroka, J. Heaton, J.T .,Gilmore, L.D., Colby, G., Roy, S., Chen, N. and De Luca, "Speech Recognition for Vocalized and Subvocal Modes of Production using Surface EMG Signals from the Neck and Face," *INTERSPEECH 2008*, Australia, 2008.

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

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

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

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

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

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

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

[10] Meltzner, G.S. Deng, Y. Colby, G. and Heaton, J.T., "Signal acquisition and processing techniques for sEMG based silent speech recognition," Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE ,.4848-4851, 2011.

[11] J.S. Garofolo, L.F. Lamel, W. M. Fisher, J.G. Fiscus, D.S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus", Linguistic Data Consortium, 1993.

[12] Special Ops Hand Signals, http://www.specialoperations.com /Focus/Tactics/Hand_Signals/default.htm

[13] Most common 1000 English Phrases, http://www.englishspeak.com/english-phrases.cfm

[14] http://www.netlingo.com/acronyms.php

[15] Y. Deng, G. Colby, J. Heaton, M. Geoffrey, "Signal Processing Advances for the MUTE sEMG Based Silent Speech Recognition System", Military Communication Conference, Oct 2012, Orlando, Florida.

[16] Schultz, T. and Wand, M., "Modeling Coarticulation in EMG-based Continuous Speech Recognition", Speech Communication Journal, Dec 2009.

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

[18] Deng, Y., Patel, R., Heaton, James T., Colby, G., Gilmore, D., Cabrera, J., Roy, S. H., De Luca, C. J., Meltzner, G. S., "Disordered Speech Recognition Using Acoustic and sEMG Signals", InterSpeech 2009, Brighton, UK.

[19] Walliczek, M., Kraft, F., Jou, S.-C., Schultz, T., and Waibel, A., "Sub-word unit based non-audible speech recognition using surface electromyography," in Proc. Interspeech, Pittsburgh, PA, Sep 2006.

[20] S. Yong, G. Evermann, (2006). *The HTK Book*.

[21] 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", in Proc. IEEE ASRU, December 2011.

[22] Schultz, T., Wand, M., Modeling Coarticulation in EMG-based Continuous Speech Recognition, Speech Communication (2009), doi: 10.1016/j.specom.2009.12.002

[23] Kumar,N., and Andreou, A.G., "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition", Speech Communication, 26(4):283-297, 1998.

[24] Gopinath, R., "Maximum likelihood modeling with Gaussian distributions for classification", in Proc. IEEE ICASSP, vol. 2, 1998, pp. 661- 664.

[25] Povey, D., Burget, L., et al., "The subspace Gaussian mixture model: A structured model for speech recognition", Computer Speech & Language, vol. 25, no. 2, pp. 404-439, April 2011.

[26] G. Colby , T. Heaton, L. D. Gilmore, J. Sroka, Y. Deng, J. Cabrera, S. Roy, C. J. De Luca, G. S. Meltzner, "Sensor subset selection for surface electromyograpy based speech recognition", *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing (ICASSP)*, 2009, Taipei, Taiwan.