Disordered Speech Recognition Using Acoustic and sEMG Signals

Yunbin Deng[1], Rupal Patel[2], James T. Heaton[4], Glen Colby[1], L. Donald Gilmore[3], Joao Cabrera[1], Serge H. Roy[3], Carlo J. De Luca[3], Geoffrey S. Meltzner[1]

[1] BAE Systems – Advanced Information Technologies, Burlington, MA, USA
[2]Communication Analysis & Design Lab, Northeastern University
[3]Delsys, Inc., Boston, MA, USA
[4]Center for Laryngeal Surgery & Voice Rehabilitation, Mass. General Hospital, Boston, MA, USA
yunbin.deng, glen.colby, joao.cabrera, geoffrey.meltzner@baesystems.com, r.patel@neu.edu, james.heaton@mgh.harvard.edu, dgilmore, sroy@bu.edu, cjd@delsys.com

Abstract

Parallel isolated word corpora were collected from healthy speakers and individuals with speech impairment due to stroke or cerebral palsy. Surface electromyographic (sEMG) signals were collected for both vocalized and mouthed speech production modes. Pioneering work on disordered speech recognition using the acoustic signal, the sEMG signals, and their fusion are reported. Results indicate that speaker-dependent isolated-word recognition from the sEMG signals of articulator muscle groups during vocalized disordered-speech production was highly effective. However, word recognition accuracy for mouthed speech was much lower, likely related to the fact that some disordered speakers had considerable difficulty producing consistent mouthed speech. Further development of the sEMG-based speech recognition systems is needed to increase usability and robustness.

Index Terms: disordered speech recognition, sEMG speech recognition, mouthed speech, silent speech, dysarthria, sEMG.

1. Introduction

Dysarthria is a motor speech disorder that arises from neurological disease such as cerebral palsy or neurological injury due to stroke or other various traumatic brain or nerve injuries. Dysarthric speech is characterized by disruption of phonation and/or poor articulatory precision, and can often be difficult for human listeners to understand. Computer recognition of dysarthric speech is a particularly challenging problem, involving overcoming several issues including: slow rate, weak volume and reduced pitch modulation, as well as inconsistency, imprecision, distortion, elongation, insertion, and deletion of phonemes.

Research on automatic recognition of disordered speech is limited [1-6] and has exclusively focused on recognition of acoustic speech. However, surface EMG (sEMG) signals generated by the facial and neck muscles have recently been proposed as a novel modality for human speech communication and human-machine interface [7-10]. Compared with traditional acoustic speech recognition, the sEMG based speech recognition has the advantages of being immune to loud background acoustic noise, capable of silent 'speech' communication with non-audible, mouthed speech, and potentially providing an alternate human-machine interface for individuals with speech impairment [11]. Recognition of speech that does not require phonation may also enable tracheostomized or ventilator-dependent individuals who have difficulty producing voice, but are often not otherwise dysarthric, to communicate rapidly and accurately.

Despite the potential advantages, sEMG-based speech recognition research is still in its infancy. The challenges include: 1) the need for multiple sensors, 2) low signal to noise ratios, 3) the lack of normative data, and 4) a lack of understanding of the relationship between the sEMG signals and the corresponding speech signal. Although previous research has shown the potential of sEMG speech recognition in high background noise environments and for silent speech communication [7][14], the effectiveness of using sEMG for disordered speech recognition has not been experimentally verified. This paper presents a pioneering effort on dysarthric speech recognition using acoustic and sEMG signals.

2. Data Collection

2.1. Participants

Five adult native speakers of English with moderate to severe speech impairment were recruited (2 females; 3 males; mean age = 45.8 years). Four participants had speech impairment due to cerebral palsy and one individual had suffered a stroke. All participants had normal hearing, vision and cognitive functioning to complete the experimental task. A motor speech evaluation was conducted to determine the dysarthria type and severity of speech impairment. Three of the five speakers presented with spastic dysarthria while the remaining two had mixed flaccid-spastic dysarthria. Severity of speech impairment was determined using the single word subtest of the Assessment of Intelligibility of Dysarthric Speech. For each speaker, intelligibility was calculated by averaging the number of correctly identified single word productions across three unfamiliar listeners. Speech intelligibility scores ranged from 60%-92% across speakers. An additional group of eight healthy speakers (4 of each gender; mean age = 24) were also recruited to serve as controls.

2.2. Stimuli and Task

Due to the expected difficulty of disordered speech recognition and the potential for vocal fatigue, our data collection effort focused on a limited vocabulary of isolated

words consisting of 11 digits, 26 NATO alphabet labels and 4 functional words (yes, no, left and right). Each participant produced a randomized order of the word list 8 times under each speaking condition (mouthed and vocalized), with the exception of dysarthric speaker 2 who only produced six mouthed lists and speaker 5 who produced voiced and mouthed word lists 10 times.

2.3. Experimental Instrumentation

The acoustic signal was collected by a headset microphone (WH30, Shure Inc., Niles, IL) which was positioned approximately 5 cm in front of and slightly lateral to the mouth while the sEMG data were collected by 11 sEMG sensors (parallel bar configuration, DE2.1, Delsys Inc., Boston MA). The sensors were positioned on the face and neck as shown in Figure 1 to provide optimal speech-related information across 6 anatomical regions (supralabial, labial, sublabial, submental neck, midline neck, and lateral neck).

To accommodate the physical needs of the dysarthric participants, the data collection sessions were conducted in their respective homes. Participants were provided with verbal descriptions of each of the mouthed and vocalized conditions and then allowed to practice to ensure that they understood the task. Once the experimenters were confident that the task was understood, data collection was initiated. The data collection protocol for the dysarthric speakers was similar to that of the healthy speakers and is detailed in a previous publication [7]. EMG signals were band-pass filtered at 20-450 Hz, the acoustic signal was low-pass filtered at 10 kHz, and all signals were digitized at 20 kHz using a 32 channel A/D converter (NI-6259, National Instruments Co., Austin, TX) and EMGWorks data acquisition software (Delsys Inc.).

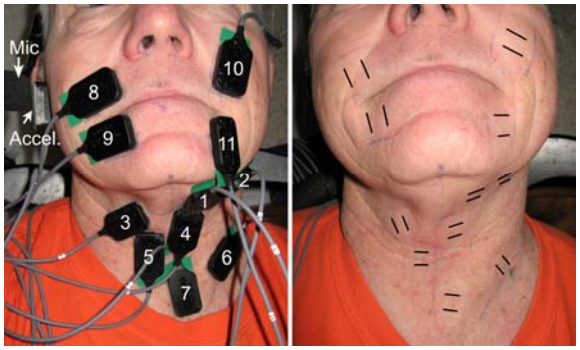

Figure 1: *The microphone and 11 sEMG sensor locations before (left) and after (right) removal. Black lines were drawn to show electrode contact locations after sensor removal.*

3. Disordered speech and sEMG signal analysis

3.1. sEMG speech activity detection

One of the main difficulties with sEMG based speech recognition is speech activity detection (SAD). We implemented an online voiced detection algorithm for sEMG based speech recognition. To achieve a rapid system response, decisions need to be made in a local short time frame; to minimize false alarms due to noisy sEMG signals, local decisions (i.e. based on a single channel) need to be adjusted globally (i.e. based on multiple channels). We only selected a subset (1, 5, 8, 9, and 11) of the 11 sEMG channels to make the SAD decision as it has been shown that high recognition performance can be obtained using only these channels [8]. Specifically, for each 50 ms window, we computed third order statistics (TOS) for each selected channel:

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

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

Figure 2 (a) shows a SAD example, with the original five sEMG channels plotted in blue, the TOS plotted in green, and the local decisions plotted in red. Note, the sEMG is plotted in units of 10 uV. In this example, sEMG activity is evident before and after the acoustic production. For comparison, a healthy speaker's acoustic signal for the same word is shown in panel (b), where it is clear that the disordered speech token has a much longer duration than does the healthy speech token.

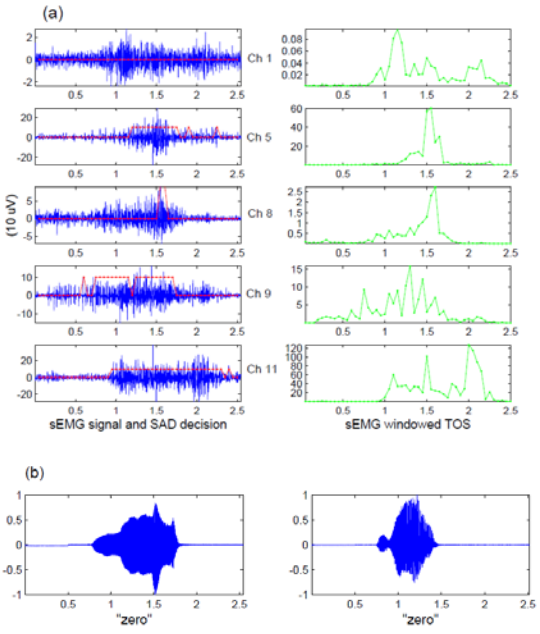

Figure 2: *(a) sEMG data for the word "zero" from a dysarthric speaker. The columns on the left are the sEMG signals, the red binary plot is the SAD decision based on TOS, which is shown on the corresponding right column. (b) Left plot is the disordered acoustic signal corresponding to the sEMG signal, on the right shows a healthy utterance for the same word.*

3.2. sEMG and acoustic feature extraction

We used Mel-Frequency Cepstral Coefficients (MFCC) with Mean and Variance Normalization (MVN) to parameterize both the acoustic and sEMG signals. Although there is no theoretical basis to use MFCCs to parameterize sEMG signals (as they are meant to approximate human auditory perceptual response to *acoustic* signals) they are a reasonable feature choice because 1) the speech-related sEMG signals and acoustic spectra both show quasi-stationary characteristics; 2) the cepstral normalization technique also helps to reduce amplitude variation of sEMG signals and the varying channel effects of sEMG sensors, and 3) our previous work has demonstrated that among a set of different candidate parameterization schemes, MFCCs produced the highest recognition rates for sEMG based speech recognition [7].

Compared with acoustic speech signals, the sEMG signals exhibit slower changes and less fine structure, thus necessitating a lower sampling rate, a slower frame rate and fewer filters in the Mel-scale filter bank. We used a root compression, $x^{0.1}$ (where x represents the output of the Mel-scale filter bank), before applying a discrete cosine transform in sEMG MFCC feature extraction. The root compression was shown to be more robust than log compression in noisy speech recognition tasks [12]. Our experiment on speaker dependent sEMG speech recognition showed similar improvement. We found that using delta-delta sEMG features showed no additional performance gain, probably due to the relative slow-varying nature of the sEMG signals. Thus, the acoustic speech signals were parameterized with a 39 dimensional MFCC feature vector, while the sEMG signals were parameterized with a 154 dimensional MFCC feature vector, consisting of concatenated features from all 11 channels. The detailed MFCC parameterization for the acoustic speech and sEMG signals is summarized in Table 1, where *E* stands for energy and *Cep* stands for cepstral coefficients.

	Acoustic Speech	sEMG
Sample rate (kHz)	16	3
Window size (ms)	25	50
Frame rate (frames/second)	100	40
Number of mel-scale filters	24	15
Features type	*E*+12 *Cep*	*E*+6 *Cep*
Delta feature	Δ, ΔΔ	Δ
Nonlinear compression	Log	Root
Total dimension	39	154

Table 1. *Summary of Acoustic and sEMG Parameterization*

4. Disordered speech recognition

4.1. Disordered speech recognition by acoustic signals alone

4.1.1. Speaker-independent acoustic model trained on healthy speech

The first recognition experiment tested the effectiveness of speaker-independent recognition for dysarthric speakers. The training data consisted of speech utterances collected from 8 healthy speakers. The same 41 words were collected, with six tokens each for each healthy speaker. A 10-state left-to-right HMM was adopted for word modeling. The testing set consisted of all tokens from the dysarthric speakers. The Hidden Markov Model Took Kit (HTK) was used to implement the HMMs in this experiment [13]. There was substantial variation in recognition accuracy among participants (see Table 2, row 1), which corresponded with dysarthria severity. The average recognition accuracy was quite low. For a similar task, healthy speakers could have achieved a much higher accuracy (>95%) [15]. We concluded that a speaker independent speech recognition system trained on healthy speakers would not be adequate for most dysarthric speakers.

Model Conditions						Recognition Accuracy (%) by Participant					
Independent	Dependent	Acoustic	sEMG	Mouthed	Vocalized	S1	S2	S3	S4	S5	Mean
x		x			x	24	58	61	27	99	**54**
	x	x			x	94	81	100	98	98	**94**
	x		x		x	81	81	100	67	96	**85**
	x		x	x		40	71	65	46	88	**62**
	x	x	x		x	96	93	100	90	99	**96**

Table 2: *Recognition accuracy for all five recognition model conditions. Note that in all conditions but the speaker independent condition, the models were trained on dysarthric speech.*

4.1.2. Speaker-dependent acoustic model trained on disordered speech

Due to the low recognition accuracy of the speaker-independent system, we generated speaker dependent models for the dysarthric speaker database. In this experiment, we used speaker-specific data to train models for each dysarthric speaker. The training data consisted of 6 tokens for each word, while the test data consisted of 2 tokens for each word, as was used in the previous test task. Although accuracy improved considerably with the speaker dependant model (see Table 2, row 2), recognition of some speakers (e.g. S2) continued to be problematic.

4.2. Disordered speech recognition using sEMG signals

Although individuals with speech impairment have difficulty producing highly consistent acoustic signals, it is possible that speech production cues may still be embedded in their facial and neck muscle activity. In this study we attempted to perform speaker-dependent speech recognition experiments using the features derived from sEMG signals. Speaker dependent models were trained and tested as was done in our previous experiments [7]. Table 2, (rows 3 and 4) summarize the results of speaker dependent sEMG speech recognition for the speech of dysarthric participants in both vocalized and mouthed speech conditions. A comparison of results in Table 2 shows that vocalized sEMG speech recognition accuracy is somewhat less effective than acoustic speech recognition. However, the mouthed speech mode has consistently yielded less accurate recognition compared to voiced speech in sEMG experiments with healthy participants [7], and this difference

in recognition between mouthed and voiced modes was even more pronounced for our dysarthric participants. During data collection, all of our dysarthric participants spontaneously commented on their difficulty producing words in the mouthed condition, finding the lack of acoustic feedback problematic and sometimes mentally exhausting. This lead to an observable inconsistency in articulation that was consistent with the relatively poor machine recognition. Additionally the lack of auditory feedback may be more detrimental for individuals with speech impairment when attempting to achieve articulatory targets. Both dysarthric and intact individuals would likely benefit from practicing mouthed speech with biofeedback of their movements (e.g. reflection of their face, sEMG signals, and real-time recognition), which will be explored in future experiments.

4.3. Multi-modal recognition

The acoustic and sEMG signals likely contain unique information relating to speech production. Previous work has demonstrated that combining these two signal streams can improve recognition accuracy especially in noisy speech conditions [14], so we combined sEMG and acoustic signals for the vocalized data set obtained from the dysarthric speakers. Because acoustic and sEMG are different types of signals and require different sampling and feature frame rates, we implemented our multi-model recognition system using a decision fusion schema. First, we trained two independent speech recognition systems (as outline above) for speech and sEMG, respectively. During testing, each recognizer generated the top five most likely hypotheses along with corresponding log likelihood score. Since the likelihood scores generated by the two systems were not on the same scale, they were normalized according to their corresponding best hypothesis:

$$nScore(i) = e^{L(i)-L(1)} \quad (2)$$

, where $L(i)$ and $nScore(i)$ are the log likelihood score and the normalized score for i^{th} best hypothesis, respectively. The best hypothesis has a value of 1 under this schema. If the same word was hypothesized by both decoders, the final score for this hypothesis was the sum of the two scores. The final scores were then sorted to generate the best final hypothesis (see Table 2, row 5), which produced notably higher average recognition rates than did either modality alone.

5. Conclusions and future work

Recognition accuracy was studied for the dysarthric speech of a small set of stroke and cerebral palsy participants using acoustic and sEMG signals during vocalized and mouthed speech modes. Our findings suggest that speaker independent acoustically based systems are inadequate for this population given large within and between speaker production variability. Custom models built for each dysarthric speaker yielded better recognition for vocalized speech using either the acoustic or sEMG signals. Additional gains in recognition accuracy were achieved by the fusion of acoustic and sEMG signals in the vocalized mode. For mouthed speech, recognition based solely on sEMG signals yielded only moderate accuracy. It is likely that mouthed speech recognition accuracy may improve if the speaker has increased opportunity to practice this new oral-motor skill and if coupled with additional phonatory signals [16]. Further work on optimizing the sensor set and the recognition models is warranted.

6. Acknowledgements

This study was sponsored by the United States Defense Advanced Research Projects Agency Defense Advanced Research Projects Agency (DARPA) Information Processing Technology Office (IPTO) Program Advanced Speech Encoding, under contract W157T-08-C-P021. 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] Polur, P.D.; Miller, G.E. “Experiments with fast Fourier transform, linear predictive and cepstral coefficients in dysarthric speech recognition algorithms using hidden Markov model.” IEEE Transactions on Neural Systems and Rehabilitation Engineering, 13(4), 558-561, 2005.

[2] Kotler, A. Thomes-Stonell, N. "Effects of speech training on the accuracy of speech recognition for an individual with a speech impairment, " Journal of Augmentative and Alternative Communication, 12: 71-80, 1997

[3] Green, P., Carmichael, J., Hatzis, A., Enderby, P., Hawley, M., Mark, P. “Automatic speech recognition with sparse training data for dysarthric speakers” European Conference on Speech Communication and Technology, 2003.

[4] Chen F. and Kostov, A. “Optimization of dysarthric speech recognition,” in IEEE EMBS Conf., 4, 1436–1439, 1997.

[5] Noyes, J. M. and Frankish, C.R., “Speech recognition technology for individuals with disabilities,” Augmentative Alternative Commun. 8, 297–303, 1992

[6] Deller Jr,, J.R., Hsu, D., and Ferrier, L.J, “On the use of hidden Markov modeling for recognition of dysarthric speech,” Comput. Methods Programs Biomed. 35, 125–139, 1991.

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

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

[9] Jorgensen, C., Lee, D.D., and Agabon, S. “Sub auditory speech recognition based on EMG signals,” Proc. Int. Joint Conf. Neural Netw., 4: 3128–3133, 2003.

[10] Manabe, H. and Zhang, Z. “Multi-stream HMM for EMG-based speech recognition,” Proc. 26th Ann. Int. Con. IEEE EMBS, 4389–4392, 2004.

[11] Lee, K., “EMG-Based Speech Recognition Using Hidden Markov Models With Global Control Variables”, IEEE Transactions on Biomedical Engineering, 55(3), 2008.

[12] Lim, J. “Spectral Root Homomorphic Deconvolution system,” IEEE Trans. on ASSP, 27 (3), 1979.

[13] Young, S. Gunnar, E, Gales, M., The HTK Book, Version 3.4, 2006.

[14] Lee, K.S. “SNR-Adaptive Stream Weighting for Audio-MES ASR”, IEEE Trans. on Biomedical Engineering. 55(8), 2008.

[15] Furui, S. “50 Years of Progress in Speech and Speaker Recognition Research,” ECTI Transaction on computer and Information Technology, Vol.1, No.2 November 2005.

[16] DiCicco, T. & Patel, R.., “Automatic Landmark Analysis of Dysarthric Speech.” Journal of Medical Speech Language Pathology, 16(4), 213-221, 2008.