Proceedings of the Federated Conference on Computer Science and Information Systems pp. 165–168
DOI: 10.15439/2017F153
ISSN 2300-5963 ACSIS, Vol. 11

Subvocal Speech Recognition via Close-Talk Microphone and Surface Electromyogram Using Deep Learning

Mohamed S. Elmahdy and Ahmed A. Morsy

Biomedical Engineering Department, Cairo University,Giza, Egypt
m.elmahdy@ieee.org, amorsy@ieee.org

***Abstract* - Speech communication is very essential for human-human communication and human machine interaction. Current Automatic Speech Recognition (ASR) may not be suitable for quiet settings like libraries and meetings or for speech handicapped and elderly people. In this study, we present an end-to-end deep learning system for subvocal speech recognition. The proposed system utilizes a single channel surface Electromyogram (sEMG) placed diagonally across the throat alongside a close-talk microphone. The system was tested on a corpus of 20 words. The system was capable of learning the mapping functions from sound and sEMG sequences to letters and then extracting the most probable word formed by these letters. We investigated different input signals and different depth levels for the deep learning model. The proposed system achieved a Word Error Rate (WER) of 9.44, 8.44 and 9.22 for speech, speech combined with single channel sEMG, and speech with two channels of sEMG respectively.**

Index Terms - Subvocal Speech; Deep Learning; sEMG.

I. Introduction

Speech plays an important role, not only in human-human communication but also in human-machine interaction. Often, human speech takes place in harsh acoustic backgrounds with a variety of environmental sound sources, competing voices, and ambient noise. The presence of such noise makes it difficult for human speech to remain robust and clear.

After the wide popularity of smart devices and assistive technologies, Automatic Speech Recognition (ASR) became the most convenient communication tool for humans to interact with these machines [1]. Although ASR systems have achieved reasonably high accuracies compared to human capabilities [2], they still suffer from various limitations. First, they are prone to environmental noise. Second, audible speech can be very disturbing in quiet settings like libraries and meetings. Third, normal speech communication is not suitable for speech handicapped, e.g., stuttering patients. Similar challenges are faced when dealing with elderly people, caused by issues with speech pace and articulation [3].

These limitations motivate the need for the development of another strategy for how ASR works in terms of speech form, acquisition techniques, and processing algorithms. One potential alternative for vocalized speech is subvocalized speech. Subvocalization occurs, for example, when someone whispers while reading a book, talking to one's self, or murmuring. This subvocalization can be acquired using surface Electromyogram (sEMG) signals from the muscles involved in speech production. Articulators involved in speech production are located in the face and neck area [4]. sEMG signals can thus be used to substitute or at least augment traditional vocalized signals.

While it is already showing great promise, the field of subvocalized speech recognition is fairly recent and not mature compared to vocalized speech recognition. Wand et al. [5] achieved a 34.7% word error rate (WER) on 50 phrases using sEMG signals of 6 facial muscles from 6 subjects. Mendoza et al. [6] obtained a WER of 25% from a single sEMG channel but only for 6 Spanish words. Wand et al. reported a 54.7% WER on 50 phrases using 35 sEMG channels and 6 subjects [7]. Furthermore, Deng et al. achieved an 8.5% WER on 1200 words using 8 channels [8].

Researchers at Nara Institute of Science and Technology, Japan [9], investigated the use of non-audible murmur microphone fabricated in their own lab using hidden markov model for further analysis, reporting a 7.9% WER. However, the NAM microphone they used isn't available, to date, for commercial or academic purposes outside of their premises.

The ability to achieve high recognition accuracies using sEMG only has proven to be very challenging [5]-[8]. The reason can be attributed to the nature of the sEMG signal, which is highly variant from subject to subject depending on the muscle strength and gender. Most of the reported research uses facial muscles to capture speech signals [6]-[8]. While giving better accuracy, this placement isn't user friendly and may not lend itself to practical implementations. Results reported in the literature used hand crafted features, heuristically chosen based on experience and visual inspection of data.

This research introduces preliminary results for a multimodal end-to-end subvocal speech recognition system using a commercially available, low cost close-talk microphone and a single channel of sEMG signal acquired from the throat area. The proposed system uses deep learning algorithms for automatic feature extraction and classification.

II. Materials and Methods

A. Corpus Design

We built an English corpus of twenty words. These words were selected to match the following criteria: (1) letters comprising the words must represent the English letters in as uniform distribution as possible, as shown in Fig.1; (2) they should be of different lengths; and (3) the similarity between words measured by Levenshtein distance [10] must be qualitatively variable, as demonstrated qualitatively in Fig.2. This limited vocabulary set can be used later for controlling machines or enabling the performance of various daily activities.

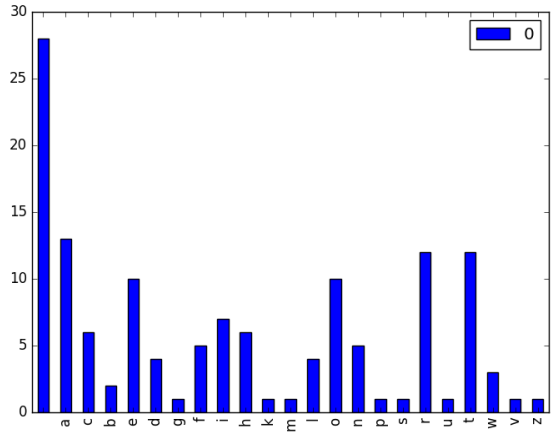

Figure 1. Distribution of letters across the corpus words

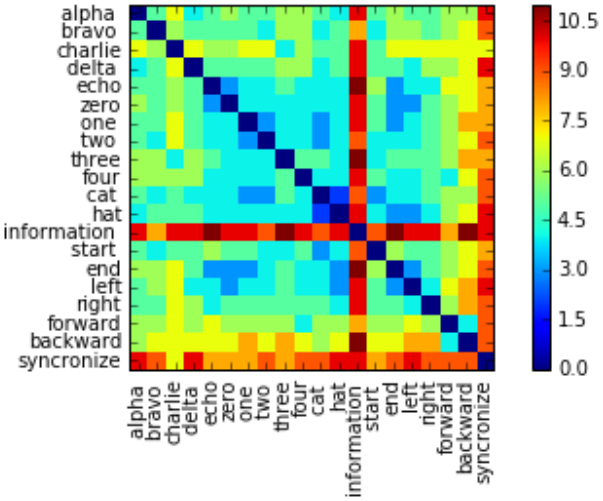

Figure 2. Levenshtein distance between words

B. *Subjects*

Ten healthy subjects participated in this experiment, five males and five females, with average age of 22 ± 2 years. All the subjects are not Native American speakers. The experiment was conducted in a lab controlled environment.

C. *Experiment Protocol and Data Labelling*

Each subject was presented with 150 slides, each containing a single phrase. Subjects were asked to subvocalize the phrase within 6 seconds of its appearance on the screen, then relax for swallowing and breathing for 10 seconds, and so on. Fig. 3 illustrates the experimental sequence and timing. The experimental setup and connections are shown in Fig. 5. Each subject gave output of 150 records, out of which 100 records were used for training and 50 records for testing, for both modalities (microphone and sEMG).

D. *Signal Acquisition*

As described above, both sEMG, to capture the electrical activities of the muscles responsible for sound production, and a close-talk microphone, to capture the articulation effect (vibration or sound), were used.

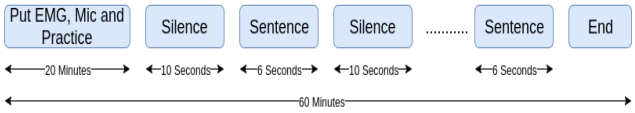

Figure 3. Timing diagram showing the experiment protocol

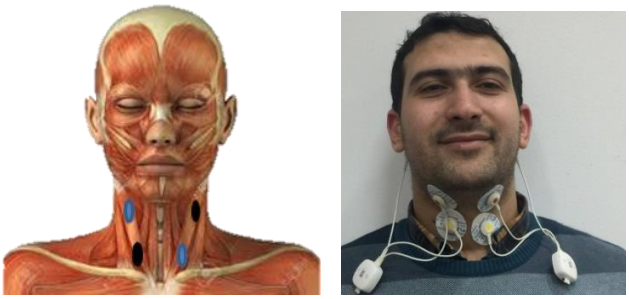

Figure 4. sEMG electrode placement

Figure 5. Experiment setup

1) Surface Electromyogram (sEMG)

A wireless sEMG from Mega Electronics Ltd was used. The sampling rate was 1 KHz using 16 bit ADC. Two electrodes were placed diagonally around the throat as shown in Fig. 4.

2) Close-Talk Microphone

A Koss CS100 close talk microphone was used. This microphone has a noise reduction filter and has a sensitivity range of -36 dB ± 3dB per 1V / 1 KHz. The microphone is placed 2 cm from the subject mouth to capture the murmurs. For recording this signal, we used freely available software named Audacity [11].

E. *Sound Pressure Level Quantification*

In order to make sure that all subjects follow the same level of subvocalization, there was a need to quantify this level numerically. We used an iPhone with an application named SPLnFFT, whose accuracy was proven by [12].

Subjects were asked to subvocalize the sentences appearing on the screen within a range of 12 ± 2 dB and were trained for 10 minutes prior to starting each recording session to help them meet this requirement.

F. *Short Time Fourier Transform (STFT)*

Input signals were converted from time domain to frequency domain through Short Time Fourier Transform (STFT) to obtain a spectrogram, which was fed to deep learning model.

G. *Deep Learning*

Conventional ASR systems consist of many complex building blocks: pre-processing, feature extraction, and building an acoustic model using Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) [13]. At the final stage, a language model is used to constrain the predicted subscription in order to follow the context of the speech as shown in Fig. 6.

Using sEMG instead of speech signals means that there is no well-defined building blocks akin to phonemes in traditional audible speech processing. Fig. 7 shows an end-to-end deep learning model as presented in [14].

H. *Spatial Convolutional Layer*

Spatial convolution layer performs the traditional convolution operation as shown in Eq. (1). Convolutional operation works to find the most similar pattern to the filter in the underlying image [15].

$$(g*f)(x,y) = \sum_{(a,b)\in A} g(a,b) f(x-a, y-b) \qquad (1)$$

I. *Bi-Directional Recurrent Neural Network (BRNN)*

RNN was developed to make use of the sequential information. In conventional neural networks, it is assumed that all the inputs and outs are independent, i.e., the input at a certain time is independent of other inputs and the same for the output.

However, for a sequential signal like sEMG and speech, this is not true. That's because each sound or letter in the previous frame affects the prediction of the sound in the next frame [16]. Also, future frames could enhance and fine-tune the prediction of earlier frames. This is the main reason for choosing bidirectional RNN instead of RNN.

J. *Connectionist Temporal Classification (CTC)*

The goal from an ASR system is to transfer any sequence of sounds into sequence of letters or phonemes. Traditional classification algorithms require that both inputs and outputs are aligned, which is not the case in most ASR problems.

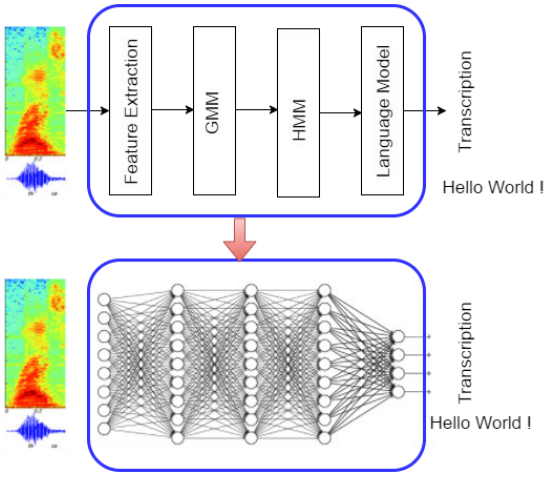

Figure 6. Building blocks for traditional ASR system versus end-to-end ASR

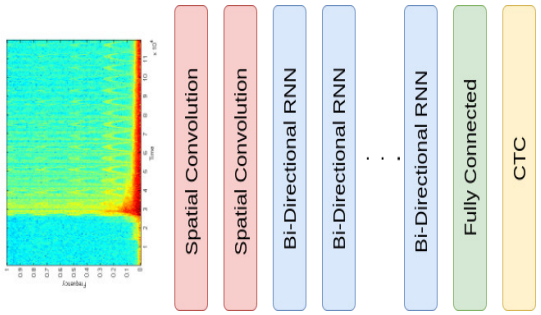

Figure 7. Proposed end-to-end deep learning model

The CTC layer generates a probability distribution at each time step of the input sequence instead of generating labels. These probabilities are then decoded into maximum likelihood labels. Finally, an objective function converts these maximum likelihood labels into the corresponding desired labels [17].

All computations were done on a desktop PC with GTX 960 TI and 6 GB GPU ram. Torch platform was used for deep learning implementation.

III. Results

This section highlights the system performance using different input signals and different number of RNN layers. For acoustic input only from the close-talk microphone, we obtained a WER of 16%, 11.33%, 9.44%, 12.56%, and 10.44% for different numbers of RNN layers as shown in Table 1.

For concatenating acoustic data and sEMG data from channel #1, we achieved a WER of 16.89%, 10.39%, 8.44%, 11.44%, and 9.83% using different RNN layers as illustrated in Table 2.

Combining acoustic data, sEMG from channel #1, and sEMG from channel #2 resulted in WER of 54.17%, 10.61%, 9.22%, 11.33, and 10.44% as shown in Table 3.

Table 1. Results for speech input

#RNN Layers	WER	CER	Time (minutes)
1	16	2.7	18.26
2	11.33	2.43	24.23
3	9.44	2	32
4	12.56	3.04	58
5	10.44	2.45	68.7

Table 2. Results for speech input + sEMG from channel 1

#RNN Layers	WER	CER	Time (minutes)
1	16.89	2.83	26
2	10.39	2.33	39
3	8.44	1.91	47
4	11.44	2.21	60
5	9.83	1.89	72.5

Table 3. Results for speech input + sEMG from channel 1 & 2

#RNN Layers	WER	CER	Time (minutes)
1	54.17	17.4	18.85
2	10.61	2.25	27
3	9.22	2.07	35.9
4	10.11	1.93	64.5
5	11.31	2.14	81.3

IV. Discussion

In this study, we investigated the performance of an end-to-end subvocal speech recognition system using a wireless sEMG system and a close-talk microphone. The performance criteria were Word Error Rate (WER) and Character Error Rate (CER). We studied the performance of the system for acoustic signal only and acoustic signal combined with sEMG from the throat muscles. The depth of the deep network model was examined in search of the optimum number of bidirectional RNN.

For the input signal being acoustic data only, we found that the performance of the system increases by increasing the number of RNN layers till a peak of 9.44% WER then it decreases by a factor 3.12% then increased by 2.12%. This sudden change in performance is likely to be caused by overfitting. After increasing the number of layers, the model starts to experience an overfitting due to the increase in the number of parameters.

When feeding the network with acoustic data concatenated with sEMG signal from the throat muscle, the performance of the system has was boosted to achieve a WER of 8.44% with an increase of 1% from acoustic signal only. This increase in performance was expected because the microphone is unlikely to catch all the information from the audio in the subvocalization mode while sEMG can capture additional information. We notice that results in Table 2 almost follow the same pattern as Table 1. Three RNN layers is the turning point for the system. After 3 layers the system experiences an overfitting problem.

For the final experiment, we fed the network with a composition of three signals: acoustic, sEMG from channel #1 and sEMG from channel #2. The best WER was 9.22% at 3 RNN layers. The performance drop illustrate that channel #2 is a noisy channel and doesn't add much information.

The timing performance for different model structures and input signals is reported in Tables 1, 2 and 3. The training time was increased when the depth of the network was increased, due to the increase in the number of parameters that need to be optimized and settled.

Comparatively, the proposed algorithm has a better performance compared to Deng et al. [8] and Wand et al. [5] in terms of WER. In contrast with literature of subvocal speech recognition, the system doesn't depend on hand crafted features or traditional building blocks for ASR. In addition, the proposed algorithm demonstrates the efficacy of a single channel sEMG combined with a close-talk microphone.

V. Conclusion

An end-to-end deep learning system for subvocal speech recognition using a close-talk microphone and a single channel wireless sEMG was presented. The proposed system used a mix of convolutional neural layers and bidirectional RNN in addition to a CTC layer as the objective layer. We studied the effect of different input signals and different numbers of RNN layers on system performance. The proposed system achieved a Word Error Rate of 9.44, 8.44 and 9.22 for acoustic, acoustic combined with a single channel sEMG, and acoustic with two channels of sEMG, respectively.

References

[1] S. Tomko, T. K. Harris, A. Toth, J. Sanders, A. Rudnicky, and R. Rosenfeld, "Towards efficient human machine speech communication," *ACM Trans. Speech Lang. Process.*, vol. 2, no. 1, pp. 1–27, Feb. 2005.

[2] W. Xiong *et al.*, "Achieving Human Parity in Conversational Speech Recognition," 2016.

[3] F. Aman, M. Vacher, S. Rossato, and F. Portet, "Analysing the Performance of Automatic Speech Recognition for Ageing Voice: Does it Correlate with Dependency Level?," pp. 9–15, 2013.

[4] S. Jou, S. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Waibel A: Towards continuous speech recognition using surface electromyography," *Proc. INTERSPEECH - ICSLP*, pp. 17--21.

[5] M. Wand, M. Janke, and T. Schultz, "Tackling Speaking Mode Varieties in EMG-Based Speech Recognition," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 10, pp. 2515–2526, Oct. 2014.

[6] L. E. Mendoza, J. Peña, and J. L. Ramón Valencia, "Electro-myographic patterns of sub-vocal Speech: Records and classification," *Rev. Tecnol.*, vol. 12, no. 2, Dec. 2015.

[7] M. Wand, C. Schulte, M. Janke, and T. Schultz, "Array-based Electromyographic Silent Speech Interface," in *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing*, 2013, pp. 89–96.

[8] Y. Deng, G. Colby, J. T. Heaton, and G. S. Meltzner, "Signal processing advances for the MUTE sEMG-based silent speech recognition system," in *MILCOM 2012 - 2012 IEEE Military Communications Conference*, 2012, pp. 1–6.

[9] Panikos Heracleous, Yoshitaka Nakajima, et al. "audible (normal) speech and inaudible murmur recognition using nam microphone", Signal Processing Conference, 2004.

[10] L. Yujian and L. Bo, "A Normalized Levenshtein Distance Metric," *IEEE Trans. Pattern Anal.*, vol.29, pp. 1091–1095, Jun. 2007.

[11] "Audacity® | Free, open source, cross-platform audio software for multi-track recording and editing." [Online]. Available: http://www.audacityteam.org/. [Accessed: 08-May-2017].

[12] D. P. Robinson and J. Tingay, "Comparative study of the performance of smartphone-based sound level meter apps, with and without the application of a ½ " IEC-61094-4 working standard microphone, to IEC-61672 standard metering equipment in the detection of various problematic workplace noise environments."

[13] J.-P. Haton, "Automatic Speech Recognition: A Review," in *Enterprise Information Systems V*, Dordrecht: Kluwer Academic Publishers, 2004, pp. 6–11.

[14] D. Amodei *et al.*, "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," Dec. 2015.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Proceedings of the 25th International Conference on Neural Information Processing Systems*. Curran Associates Inc., pp. 1097–1105, 2012.

[16] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, "Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks," Jan. 2017.

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 2006, pp. 369–376.