CROSS-SPEAKER SILENT-SPEECH COMMAND WORD RECOGNITION USING ELECTRO-OPTICAL STOMATOGRAPHY

Simon Stone and Peter Birkholz

Institut of Acoustics and Speech Communication, Technische Universität Dresden, Germany

ABSTRACT

Speech recognition based on articulatory movements instead of the acoustic signal is of growing interest in the community. In this work, we present the results of a study using a novel measurement technology called Electro-Optical Stomatography to capture speech movements and use the acquired data to recognize a number of command words. The performance of the recognition system was evaluated using two vocabularies (one with 30 and one with 10 words) and four speakers. The speaker-dependent results were up to the state-of-the-art with average word accuracies of 97 % to 99.5 %, while the speaker-independent results exceeded it with average word accuracies of approx. 56 % to 62 %.

***Index Terms*—** Silent-Speech, Electro-Optical Stomatography, SSI, EOS

1. INTRODUCTION AND RELATED WORK

The rise of speech-based Human-Machine Interaction has been greatly accelerated by recent developments in artifical neural networks and deep learning and today speech-recognition-powered interfaces are omnipresent. However, speech as an input modality has two major disadvantages: it comes with inherent privacy concerns (because it can be easily overheard by bystanders) and it may not be comfortable or even available to many users, e.g., laryngectomized cancer patients or elderly people. Still, the naturalness of spoken communication is a major upside and so there have been increased research efforts into keeping the convenience of speech interfaces without the caveats mentioned above: By ignoring the acoustic signal and going straight to the source of speech, the articulatory movements of the speech organs (lips, tongue, jaw, and so on) — a silent-speech interface (SSI). A major roadblock for these efforts is to find a robust and portable technique to capture the speech movements with high precision and, especially, with a sufficient reliability and reproducibility across different subjects - an "articulatory microphone", so to speak.

Many different technologies have been considered for this task (see [1] and [2] for two thorough reviews of the field), but only two technologies have emerged as both powerful and portable enough to have a significant chance to become the "articulatory microphone" that will power silent-speech interfaces: electromyography (EMG), which measures the electric activity of the muscles participating in the speech production, and permanent-magnetic articulography (PMA), which measures the changes in a magnetic field caused by the movement of permanent magnets glued to the tongue, lips, and jaw. The first study attempting to recognize words using EMG [3] was comprised of five experiments differing by the vocabulary and subject used: The largest vocabulary was a set of 17 pseudowords, while another investigated vocabulary contained the ten English digit words ("zero", "one", "two" and so on). In the first case, each pseudoword was repeated 10 times. The numbers corpus contained 20 repetitions of each word. The overall accuracy (number of words correctly recognized divided by total number of words) was 35 % for the 17 pseudowords and approx. 65 % for the numbers. The study also investigated the accuracy of classifiers trained with data from one speaker and tested on data from the other speaker (inter-speaker dependency) and with data from the same speaker but from another recording session (inter-session dependency). The result was that the accuracy dropped from around 35 % (same speaker and session) to 30 % (same speaker, different session) to below chance level (different speaker, different session). Subsequent work in the field by the groups around Meltzner [4, 5, 6] and Schultz [7, 8, 9] greatly improved the initial results and expanded the paradigm to continuous speech recognition, but have yet to overcome the strong session dependency (and, by extension, speaker dependency): the best average cross-session accuracy achieved in [9] was 71.5 % for a vocabulary of 108 words. In a concurrent research effort using a different technology [10], a research group around Fagan started investigating the suitability of permanent-magnetic articulography (PMA) for use in an SSI. Their PMA device consisted of a wearable support structure (similar to the frame of a pair of eyeglasses) carrying six dual axis magnetic sensors. During (silent or audible) speech, the magnetic field changed at the sensor positions because of the relative movement of permanent magnets attached to the center of the subject's tongue tip and

This work was funded by the German Federal Ministry of Education and Research, reference number 13GW0101B and the German Federal Ministry for Economic Affairs and Energy, reference number ZF4443004BZ8.

to the upper and lower lips. Their first study consisted of two experiments: one experiment using a vocabulary of 13 isolated phonemes and one experiment using a vocabulary consisting of 9 words. Each word or phoneme was spoken 10 times by a single subject. The recognition accuracy with this setup was 94 % for the 13 phonemes and 97 % for the 9 words, using a nearest-neighbor classifier. In [11], the authors used the same measurement device and classifier, but different vocabularies: one vocabulary consisting of the 10 English digit words "zero" to "nine" (the numbers set) and one vocabulary additionally consisting of 47 other words (the words set) chosen to cover a wide range of phones. Using ten repetitions of the numbers set and five repetitions of the words set, the nearest-neighbor classifier was evaluated using leave-one-out cross-validation for three speakers. The recognition rates ranged from 82 % to 100 % for the numbers set and from 76 % to 99 % for the words set, depending on the speaker. In [12, 13], they replaced the classifier with statistical sequence modeling using Hidden Markov Models (HMM) and thus achieved a leave-one-out cross-validated word accuracy of 92 % to 98.8 % on the words set for a single speaker, depending on the signal condition used. Going a little further in [13] they also performed a digit sequence recognition experiment using HMMs and achieved a sequence accuracy of 61.1 % to 81.7 %. While these results were all obtained using only a single speaker, in [14] the same general setup was used to train and test models for three speakers individually (again using only data from the same speaker for training and testing). The results for the word accuracy ranged from 82.72 % to 90.97 % and the sequence accuracy from 74.89 % to 86.76 %. So far, all of their studies have focused on intra-speaker evaluations, while the inter-speaker performance remains yet to be evaluated.

A third and yet underexplored technology specifically developed for the capture of tongue and lip movements is called electro-optical stomatography (EOS) [15, 16, 17]. It measures both the palato-lingual contact pattern (using electrical contact sensors) and the distance between the tongue and palate, as well as the lip opening and protrusion (using optical sensors). In a previous study [18], we evaluated a command word recognizer using EOS data on a dataset similar to the sets employed in the other early studies of the state-of-the-art technologies, using just a single speaker, and achieved an accuracy of 52 %. In this study, we greatly improved the results and expand the scope of the study by using further developed measurement hardware, a more sophisticated classification scheme, and by investigating the inter-individual performance differences in speaker-specific and cross-speaker paradigms.

2. THE DATASET

Our system uses a pseudopalate with 32 contact sensors, 5 laser-optical distance sensors (each consisting of a laser diode and a phototransistor) along the midsagittal line and two additional optical lip sensors to capture the lip opening and protrusion, respectively. It gathers data at a frame rate of 100 Hz. An example EOS palate is shown in Figure 1a. This study used four speakers (all male, age 30-41) and the same 30 most common German words as in the previous study [18] and additionally the ten German digit words for the digits 0 to 9 (similar to the setup in [11], see Table 1). Each group of words was repeated 10 times for a total of 300 instances in the frequent words data set and 100 instances in the numbers data set for each speaker. The data was collected using a custom PC software, a Plantronics Blackwire C720 M stereo headset (for reference audio) and an EOS device with the internal version number 3.2. The recordings were made in a quiet office environment. The speakers were prompted to read a carrier word (the German indefinite article "eine" - /ˈaɛ̯nə/) followed by the word of interest. The schwa /ə/ at the end of the carrier word ensured a neutral vocal tract configura-

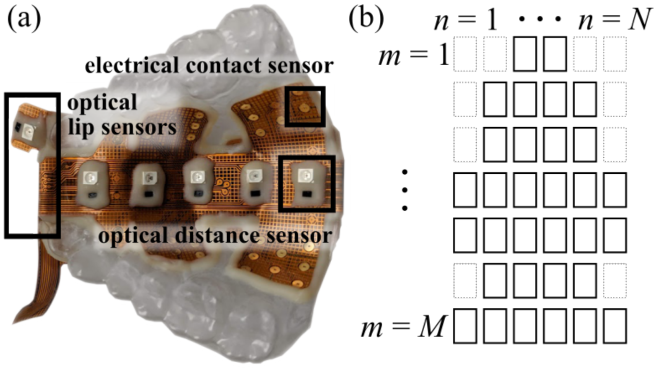

Fig. 1. **(a)** Example pseudopalate with sensors for EOS measurements. One optical lip sensor faced the lower lip (to measure lip opening) and one was placed on the incisors facing the inside of the upper lip to measure lip protrusion (occluded in the figure). Each pseudopalate was fitted to each individual speaker. **(b)** Matrix for calculating the contact pattern factors: solid boxes represent the contact sensors and can adopt the values 0 (no contact) or 1 (contact), dashed boxes represent always-off pseudo-sensors inserted to obtain an $M \times N$ matrix shape.

Noun		Adjective		Verb		Digit	
Jahr	jaːᴿ	neu	nɔœ̯	werden	vˈeːᵄdn̩	Null	nʊl
Uhr	uːᵄ	andere	ˈandəʁə	haben	hˈaːbm	Eins	aɛ̯ns
Prozent	pʁ̥otsˈɛnt	groß	gʁoːs	sein	zaɛ̯n	Zwei	tsv̥aɛ̯
Million	mɪli̯on	erste	ˈeːᵄstə	können	kˈœnən	Drei	dʁaɛ̯
Euro	ˈɔœ̯ʁoː	viel	fiːl	müssen	mˈʏsn̩	Vier	fiːᵄ
Zeit	tsaɛ̯t	deutsch	dɔœ̯tʃ	sollen	zɔln	Fünf	fʏnf
Tag	taːk	gut	guːt	sagen	zˈaːgŋ̍	Sechs	zɛks
Frau	fʁ̥aɔ̯	weit	vaɛ̯t	geben	gˈeːbm̩	Sieben	zˈiːbm̩
Mensch	mɛnʃ	klein	klaɛ̯n	kommen	kˈɔmən	Acht	axt
Mann	man	eigen	ˈaɛ̯g ŋ̍	wollen	vɔln	Neun	nɔœ̯n

Table 1. Standard pronunciation of the words used in the study (according to [19])

tion at the beginning of the word of interest. The words were produced in a natural way, i.e., with phonation and at an unregulated speaking rate of each speaker's individual choice. The EOS data was manually segmented so that each training sequence only contained data from the actual articulation of the target word (and not from the carrier word or from the neutral vocal tract configuration between items). A sequence of feature vectors was created from the segmented data. Each feature vector consisted of the ADC data of the 2 lip sensors, 5 distance sensor values, and 3 factors describing the contact pattern for a total of 10 features per vector. The distance sensor values were (depending on the hyperparameter setting) either raw ADC values or converted to mm using the calibration scheme described in [16], the latter potentially reducing the in-session variance of the measurements. The contact pattern factors were calculated to reduce the dimensionality of the feature vectors, as the raw contact pattern would introduce 32 binary features instead. The chosen factors were the normalized sum of the activity s, the center of gravity c, and the laterality measure l. The sum of the activity s was defined as:

$$s = \frac{1}{K} \sum_{m=1}^{M} \sum_{n=1}^{N} x(m,n) \tag{1}$$

with K being the total number of contact sensors (32), M, N being the number of rows and columns of the contact pattern matrix (7 and 6, respectively) and $x(m,n)$ being the binary contact sensor value at the position (m,n) in the pattern. Using the same naming conventions, the center of gravity c calculation was:

$$c = 1 - \frac{\sum\limits_{m=1}^{M} \sum\limits_{n=1}^{N} (m-0.5)\, x(m,n)}{M \cdot \sum\limits_{m=1}^{M} \sum\limits_{n=1}^{N} x(m,n)}, \tag{2}$$

and finally, the laterality measure l was given by:

$$l = \frac{\sum\limits_{m=1}^{M} \sum\limits_{n=1}^{N} |n - \frac{N+1}{2}| \cdot x(m,n)}{\frac{N}{2} \sum\limits_{m=1}^{M} \sum\limits_{n=1}^{N} x(m,n)}. \tag{3}$$

3. TRAINING

A bidirectional long short-term memory (BLSTM) network was trained to recognize the command words and validated using MATLAB 2018b. The network was trained for both data sets (numbers and frequent words) independently. Due to the large number of hyperparameters of BLSTM networks, some of the hyperparameters were set to reasonable, fixed values: the number of hidden layers was set to 1 because of the small number of training data. The gradient was clipped at 1, which is common practice to avoid the exploding gradient problem. The number of training epochs was set sufficiently large (300) but at the same time early stopping was used, so the training never actually timed out but was always stopped due to a diverging validation loss. The validation frequency for early stopping was 10 iterations and the validation patience was 5. These values were empirically determined by manual examination of the training progress with various settings. The learning rate was set to a constant value of 0.01 and the size of the mini-batches was aligned with the size of the respective vocabulary (i.e., 10 in the case of the numbers data set and 30 in the case of the frequent words data set). The number of neurons N in the hidden layer, the dropout ratio δ, and the choice of the data format (raw ADC values or converted to mm) were subject to Bayes optimization with 30 evaluations of the cost function. The search space for these was the integer interval between 100 and 256 for N and the continuous interval between 0.2 and 0.9 for δ. All other hyperparameters and options were set to the default values suggested by MATLAB. A summary of the hyperparameter settings is shown in Table 2.

Hyperparameter	**Evaluated values/ranges**
Number of hidden layers	1
Number of neurons N	[100, 256]
Dropout δ	[0.2, 0.9]
Gradient threshold	1
Max. number of epochs	300
Validation frequency	10
Validation patience	5
Learning rate	0.01
Mini-batch size	10 (numbers), 30 (frequent words)
Data format	{ADC, mm}

Table 2. Hyperparameter settings of the BLSTM networks. The optimal hyperparameter combination was found using Bayes optimization with 30 evaluations of the cost function.

4. RESULTS

Two different evaluation paradigms were used: a speaker-dependent evaluation and a cross-speaker evaluation.

4.1. Intra-speaker validation

In this paradigm, the data sets recorded with each of the four speakers were used independently to train the BLSTM network. Since both data sets contained 10 repetitions of each item, one instance of each item was excluded from training and used for evaluation while the other 9 instances were used for training the network. This strategy is a special kind of leave-one-out cross-validation or non-randomly partitioned 10-fold cross-validation and was chosen to keep the number of models to train low while at the same time giving a fair estimation of the accuracy of the prediction on unseen data. The accuracy on the evaluation set was measured by predicting the label for each instance and determining the percentage of correct predictions. This procedure was repeated until every instance was part of the evaluation set once. The results using the optimal hyperparameters (see previous section) are given in Table 3.

4.2. Inter-speaker validation

Articulatory data is generally highly specific to each individual. In the EOS data, the main differences likely originate in the different sensor positions relative to each subject's anatomy of their anterior mouth cavity. While the sensors have the same relative positions to one another because they are mounted to flexible circuit boards of the same layout, the incisor geometry and curvature of the hard palate is different for every subject. Therefore, the optical axes of the optical sensors were at different angles for different subjects and the contact sensors ended up in different areas of the hard palate. To quantify the impact of this inter-speaker variability in the sensor data, another set of four BLSTM networks was trained using the data from three of the four subjects for training and the data of the fourth subject for validation so that every subject's data was used for validation once (leave-one-speaker-out cross-validation). The hyperparameter tuning followed the same procedure as described in section 4.1 and used the same search space. The results are shown in Table 4.

Numbers set

Subject	**1**	**2**	**3**	**4**	**Average**
Hyper-parameters	$N = 151$, $\delta = 0.31$, [mm]	$N = 132$, $\delta = 0.14$, [ADC]	$N = 133$, $\delta = 0.88$, [ADC]	$N = 123$, $\delta = 0.34$, [mm]	
Accuracy	100 %	99 %	99 %	100 %	**99.5 %**

Frequent words set

Subject	**1**	**2**	**3**	**4**	**Average**
Hyper-parameters	$N = 132$, $\delta = 0.84$, [mm]	$N = 215$, $\delta = 0.39$, [mm]	$N = 202$, $\delta = 0.16$, [ADC]	$N = 248$, $\delta = 0.67$, [ADC]	
Accuracy	96.67 %	96.67 %	98.33 %	96.33 %	**97 %**

Table 3. Recognition accuracy in the intra-speaker evaluation on the numbers corpus (above) and the frequent words corpus (below).

5. DISCUSSION AND OUTLOOK

The results of the intra-speaker evaluation are comparable to the state-of-the-art in the field set by [11] using PMA. Their reported errors were in-sample errors, however, and swing wildly between the two subjects. The results from this study were obtained in a more systematic fashion and are more consistent across the speakers, while at the same time slightly surpassing the previous benchmarks. EOS therefore appears to capture the individual's articulation sufficiently well to discriminate between a limited set of words. It remains to be investigated if and how the accuracy decreases with increasing vocabulary size. The inter-speaker analysis identified further room for improvement. The inter-individual differences of the speakers led to a precipitous drop in accuracy from an average of 99.5 % to an average of 61.75 % on the numbers corpus and from an average of 97 % to an average of 56.17 % on the frequent words corpus. However, even for the worst evaluation speaker the accuracy was still better than any results from comparable systems, even though no additional adaptation was done. Also, the variance of the accuracy across speakers is quite high: the achieved performance ranged from 33 % all the way to 82 %, depending on the evaluation speaker (see Table 4). Nevertheless, a speaker adaptation of some sort is needed to achieve practically useful accuracy levels. This adaptation could be as simple as obtaining more training data from more speakers, or more elaborate and involve finding "alignment utterances" that map a speaker's articulatory space to a generic model speaker's space, in which the classification is subsequently performed. In both cases, more data needs to be acquired before further investigations can be pursued. In summary, the results from this study using the rather new technique EOS met or exceeded the results of earlier studies using state-of-the-art systems employing EMG and PMA at a comparable stage of their development cycle, and thus suggest the potential to surpass the state-of-the-art after further development. Also, the speaker-independent results showed that EOS appeared to be less prone to speaker variability "out-of-the-box". Until entirely non-invasive techniques for articulatory data acquisition (e.g., EMG [9] or radar-based measurements [20]) have matured further, EOS therefore appears to be a good compromise between invasiveness and reproducibility of the measurements. Future work will endeavor to walk through the same steps that the current state-of-the-art has taken (see section 1): expand the vocabulary and extend the setup to continuous speech recognition, while tackling the problem of the remaining speaker dependency by using further post-processing of the data, e.g., by using alignment utterances (see above).

Hyperparameters: $N = 117$, $\delta = 0.45$, [mm]

Evaluation speaker	**1**	**2**	**3**	**4**	**Average**
Accuracy	33 %	52 %	80 %	82 %	**61.75 %**

Hyperparameters: $N = 119$, $\delta = 0.387$, [mm]

Evaluation speaker	**1**	**2**	**3**	**4**	**Average**
Accuracy	56 %	52.33 %	63.67 %	52.67 %	**56.17 %**

Table 4. Recognition accuracy in the inter-speaker evaluation on the numbers corpus (above) frequent words corpus (below) using a different speaker for testing for each network (leave-one-speaker-out cross-validation).

6. REFERENCES

[1] Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, Jim M Gilbert, and Jonathan S Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[2] Tanja Schultz, Michael Wand, Thomas Hueber, Dean J Krusienski, Christian Herff, and Jonathan S Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, 2017.

[3] Michael S Morse and Edward M O'Brien, "Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes," *Computers in Biology and Medicine*, vol. 16, no. 6, pp. 399–410, 1986.

[4] Geoffrey S Meltzner, Glen Colby, Yunbin Deng, and James T Heaton, "Signal acquisition and processing techniques for sEMG based silent speech recognition," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 4848–4851.

[5] Yunbin Deng, James Heaton, and Geoffrey Meltzner, "Towards a practical silent speech recognition system," in *Proc. of the Interspeech*, Singapore, 01 2014, pp. 1164–1168.

[6] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, "Silent speech recognition as an alternative communication device for persons with laryngectomy," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2386–2398, Dec 2017.

[7] Michael Wand and Tanja Schultz, "Session-independent EMG-based speech recognition.," in *Biosignals*, 2011, pp. 295–300.

[8] Michael Wand and Tanja Schultz, "Towards real-life application of EMG-based speech recognition by using unsupervised adaptation," in *Proc. of the Interspeech 2014*, Singapore, 2014, pp. 1189–1193.

[9] Michael Wand, Tanja Schultz, and Jürgen Schmidhuber, "Domain-adversarial training for session independent EMG-based speech recognition.," in *Proc. of the Interspeech 2018*, Hyderabad, India, 2018, pp. 3167–3171.

[10] Michael J Fagan, Stephen R Ell, James M Gilbert, E Sarrazin, and Peter M Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Medical Engineering & Physics*, vol. 30, no. 4, pp. 419–425, 2008.

[11] James M Gilbert, Sergey I Rybchenko, Robin Hofe, Stephen R Ell, Michael J Fagan, Roger K Moore, and Phil Green, "Isolated word recognition of silent speech using magnetic implants and sensors," *Medical Engineering & Physics*, vol. 32, no. 10, pp. 1189–1197, 2010.

[12] Robin Hofe, Stephen R Ell, Michael J Fagan, James M Gilbert, Phil D Green, Roger K Moore, and Sergey I Rybchenko, "Evaluation of a silent speech interface based on magnetic sensing," in *Proc. of the Interspeech*, Makuhari, Chiba, Japan, 2010, pp. 246–249.

[13] Robin Hofe, Stephen R Ell, Michael J Fagan, James M Gilbert, Phil D Green, Roger K Moore, and Sergey I Rybchenko, "Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing," *Speech Communication*, vol. 55, no. 1, pp. 22–32, 2013.

[14] Robin Hofe, Jie Bai, Lam Aun Cheah, Stephen R Ell, James M Gilbert, Roger K Moore, and Phil D Green, "Performance of the MVOCA silent speech interface across multiple speakers," in *Proc. of the Interspeech*, 2013, pp. 1140–1143.

[15] Peter Birkholz, Philippe Dächert, and Christiane Neuschaefer-Rube, "Advances in combined electro-optical palatography," in *Proc. of the Interspeech 2012*, Portland, Oregon, USA, 2012, pp. 703–706.

[16] Simon Preuß and Peter Birkholz, "Optical sensor calibration for Electro-Optical Stomatography," in *Proc. of the Interspeech 2015*, Dresden, Germany, 2015, pp. 618–622.

[17] S. Stone and P. Birkholz, "Angle correction in optopalatographic tongue distance measurements," *IEEE Sensors Journal*, vol. 17, no. 2, pp. 459–468, Jan 2017.

[18] Simon Stone and Peter Birkholz, "Silent-speech command word recognition using electro-optical stomatography," in *Proc. of the Interspeech 2016*, San Francisco, CA, USA, 2016, pp. 2350–2351.

[19] E.M. Krech, E. Stock, U. Hirschfeld, L.C. Anders, P. Wiesinger, W. Haas, and I. Hove, *Deutsches Aussprachewörterbuch*, De Gruyter, Berlin, Germany, 2009.

[20] P. Birkholz, S. Stone, K. Wolf, and D. Plettemeier, "Non-invasive silent phoneme recognition using microwave signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2404–2411, Dec 2018.