EVALUATING OPTOPALATOGRAPHY SENSOR POSITIONS FOR SILENT COMMAND WORD RECOGNITION

Arne-Lukas Fietkau, João Menezes*, Peter Birkholz*

Institute of Acoustics and Speech Communication, Dresden University of Technology
{arne-lukas.fietkau, joao_vitor.possamai_de_menezes}@tu-dresden.de

Abstract: Optopalatography is an articulatory measurement technique with a broad range of applications, e.g. in speech therapy and Silent Speech Interfaces (SSI). This papers uses a custom-developed OPG device as SSI for the task of command word recognition and investigates the relevance of different sensing positions, namely at the lips, at midsagittal positions, and at lateral positions, for the recognition accuracy via an ablation study. A corpus consisting of 10 repetitions of 100 different words was recorded by 4 speakers and served as input data for a recurrent neural network. Three types of evaluations were carried out: single speaker, leave-one-speaker-out and multi-speaker. The mean accuracies obtained in the single speaker and leave-one-speaker-out evaluations were 81.25% and 47.93%, respectively. The multi-speaker accuracy was 90.25% and is comparable to the state-of-the-art. The ablation study results showed that the single speaker recognition using only midsagittal sensors yielded a relative decrease in accuracy of 4.92% in comparison to when all sensors were considered. In multi-speaker evaluation, on the other hand, sensing configurations using either lip or lateral sensors showed relative accuracy decreases of 2.49% and 1.94%, respectively. The more sensors are removed from the input data, the larger are the accuracy decreases, meaning all sensing positions improve recognition accuracy.

1 Introduction

Devices capable of measuring speech articulation find applications in a variety of activities, such as Silent Speech Interfaces (SSIs), speech rehabilitation and experimental phonetics. These applications are however limited by the characteristics of each device, e.g. portability, invasiveness, robustness, and temporal/local resolution [1]. For example, a measurement technique which is appropriate for laboratory phonetic measurements like electromagnetic articulography (EMA) due to its accuracy, might be inappropriate for Silent Speech Interfaces or even for field phonetic measurements, since it is not portable and demands a rigid procedure to attach its sensors [2].

Optopalatography (OPG) is one technique for measuring speech articulation whose features such as portability and robustness [3] make a viable alternative for a diverse range of applications. So far, OPG has been applied in speech therapy [4, 5] and SSIs [6, 7]. This paper presents a study on OPG-based command word recognition, which is a task commonly performed by SSIs as its generally easier than continuous speech recognition, but already enables a series of use cases, e.g. human-machine interaction.

The task of silent command word recognition, i.e., the recognition of spoken command words without the use of acoustic speech data, has been tackled using a variety of sensing

*Equal contributors

modalities. Using electroencephalography (EEG), Vorontsova et al. [8] achieved 84.5% recognition rate on a single speaker corpus of 9 words using a deep neural network composed of convolutional and recurrent layers. With surface electromyography (sEMG) sensors, Wu et al. [9] classified between 100 daily-life commands using deep convolutional neural networks with an accuracy of 90.67% in a multi-speaker setting (28 total speakers). A custom developed stepped frequency continuous wave radar system by Wagner et al. [10] was used to record a 50-word corpus from 2 speakers, which was then classified by a recurrent neural network with a mean speaker-dependent accuracy of 99.17%. More recently, non-conventional sensing modalities also showed promising results, focusing on the portability feature and developing wearable devices. Zhang et al. [11] applied acoustic sensing of the speaker's skin deformation during speech to classify between 31 isolated commands recorded by 12 speakers with 95.5% accuracy using deep convolutional neural networks. Sun et al. [12] sensed the speaker's ear canal deformation using earphones and were able to achieve 93% recognition accuracy using deep convolutional neural networks in a 50-word corpus in a multi-speaker setting (50 total speakers).

The performance of OPG-based systems in command word recognition was evaluated by our group in two previous studies. Stone and Birkholz [6] used the electro-optical stomatography (EOS) technique, which consists of a combination of electrical contact sensors and optical distance sensors mounted in a pseudopalate, to classify two corpora: one consisting of 10 digits from "Null" to "Neun" (German words for "zero" and "nine"), and one consisting of the 30 most common German words (10 adjectives, 10 nouns and 10 verbs). Both corpora were recorded by 4 native German speakers. The study performed command word recognition in a single speaker setting (accuracies between 97% and 99.5%) and in a leave-one-speaker-out setting (accuracies from 56% to 62%), that is, testing the system with data from a complete unseen speaker. The neural architecture used was a recurrent neural network composed of a single bidirectional long short-term memory (BLSTM) layer. Possamai de Menezes et al. [7] used the same OPG device as the present study and a 40-word corpus (union of both corpora used in [6]) recorded by one speaker to achieve a 98.38% command word recognition accuracy. This result was obtained with a 5-fold cross validation of a dynamic time warping (DTW)-based pattern matching algorithm.

The present study presents itself as an expansion of our previous research on OPG command word recognition to compare the performance of our newly designed OPG2023 system with former work. Additionally, an ablation study of different sensing locations in the palate is conducted with the aim of guiding the development of future versions. In this regard, the motivation is to investigate which (sub)set of the 15 optical sensors provides the optimal classification performance while keeping the device as simple as possible.

2 Method

An extended corpus of 100 words was set in order to set a more difficult task. Audio and articulatory data were recorded by 4 individuals. The articulatory data was then used to train a neural network with it.

2.1 Articulatory recording

The self-developed optopalatographic device "OPG2023" was used to record the articulation. Its sensors consist of laser-phototransistor-pairs and are measuring the reflected light from an object (here tongue and lips). Fifteen sensors are placed on one pseudo palate and are positioned as seen in Figure 1 to record important articulatory areas of the tongue and lips. The current induced by the incoming light intensity is measured with an 18 bit ADC. The dependency

between the distance and the light intensity is shown in Figure 2. OPG2023 records with a 100 Hz sampling rate and it has a mechanism for ambient light compensation. A built-in acoustic buzzer on the device can be used for synchronizing the audio and articulatory recordings.

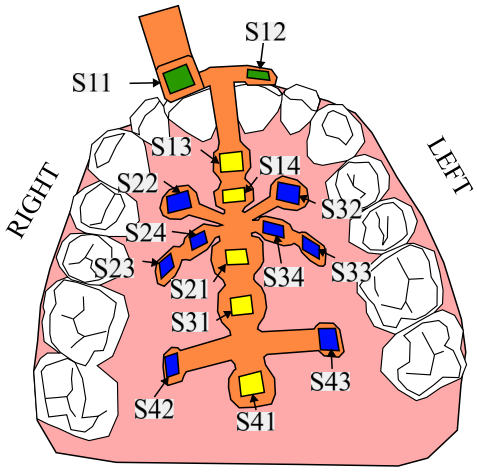

Figure 1 – Sensor positions on the artificial palate. The colour coding means: green - lip sensors, yellow - midsagittal tongue sensors, blue - lateral tongue sensors.

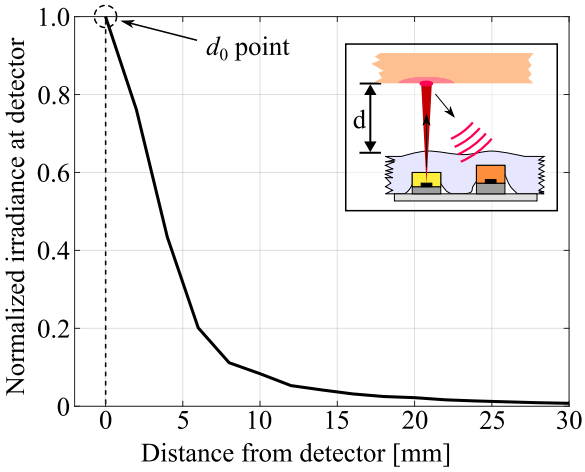

Figure 2 – Example of a tongue-distance-curve (from [13], edited).

2.2 Corpus, subjects and experimental procedure

The corpus is composed of three different sources: It contains 80 German words from two former mentioned studies [6, 10]. Additionally, 20 English words (see Table 1) were added for the potential usage in another task.

Table 1 – Set of added command words. The IPA transcriptions were taken from [14].

Robot command items							
Spot	spɒt	tongue	tʌŋ	faster	faːstə	backward	bækwəd
start	stɑːt	off	ɒf	slower	sloʊə	sneak	sniːk
stop	stɒp	shutdown	ˈʃʌtdaʊn	left	left	walk	wɔːk
hello	həˈloʊ	on	ɒn	right	raɪt	run	rʌn
manual	mænjuəl	no	noʊ	forward	fɔːwəd	sit	sɪt

Four healthy male subjects (S1: fluent German speaker, S2 - S4: native Germans) with the ages 30-45 years, who gave informed consent, were recorded in an audio studio. Everyone used their own individually manufactured OPG pseudo palate. Articulatory and audio data were recorded by using the software Articulatory Data Recorder 2 (ADR2, described in more detail in [7]). Figure 3 depicts the recording setup and Figure 4 the use of it. Each subject performed 10 repetitions of all 100 words, which were recorded in one or two sessions. The corpus was shuffled in each repetition. In total, a dataset of 4000 examples was acquired (100 words $\times$ 10 repetitions $\times$ 4 subjects). Examples of the recorded articulatory data are shown in Figure 5.

Figure 3 – Recording setup: laptop with ADR2, OPG2023 and microphone.

Figure 4 – Usage of the setup during recording.

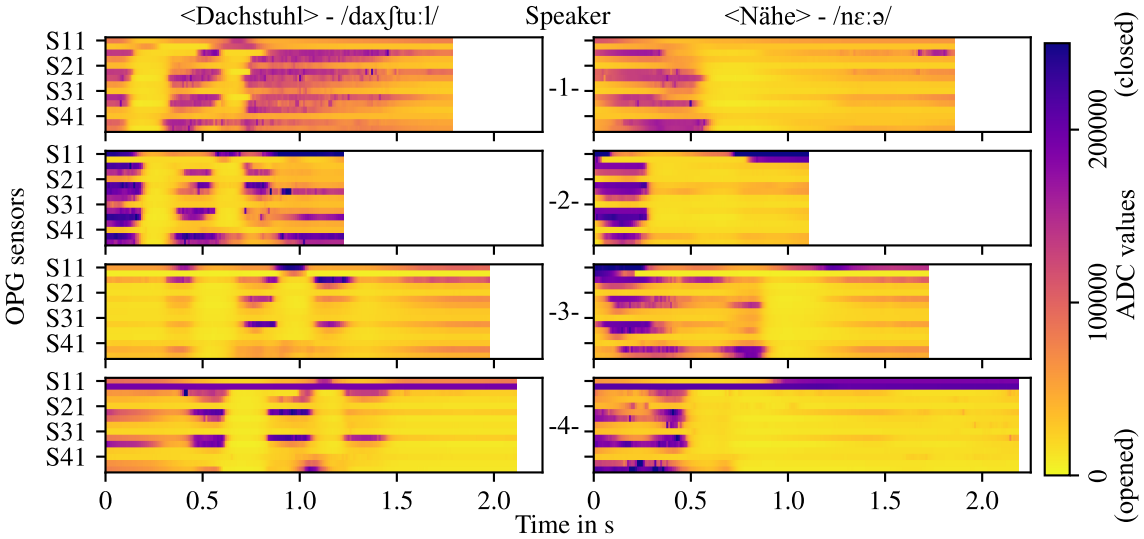

Figure 5 – Repetitions of the words "Dachstuhl" and "Nähe" by all 4 speakers represented by the ADC values of the 15 OPG sensors. High ADC values mean more proximity to the OPG artificial palate. Note that each word is realised with a similar pattern, despite coming from different speakers.

2.3 Neural network architecture

The command word recognition experiments were carried out using the raw ADC values recorded by OPG2023 as input, and the 100 words in the corpus as target labels.

For comparison reasons, the same neural architecture and training parameters as in [6] were used. The chosen architecture is composed of a BLSTM layer, followed by a fully connected and a *log softmax* layer. Table 2 presents the used hyperparameter values. For each evaluation, 30 random-search hyperparameter optimization runs were performed. Prior to training, 30 random hyperparameter sets were generated and used consistently across this study. The best-performing hyperparameters for all evaluations are available on `https://vocaltractlab.de/index.php?page=birkholz-supplements`.

We performed single speaker and leave-one-speaker-out evaluations, as in [6]. In addition, we also performed a multi-speaker evaluation, with data from all 4 speakers considered together as a single dataset. Table 3 describes the composition of training, validation and test sets for each type of evaluation. Normalisation of the OPG input data was done according to a *MinMax* strategy, compressing the signal of each sensor in the 0 to 1 range. Each sensor was normalised individually and based on the statistics (minimum and maximum values) of the training set.

Table 2 – Description of the hyperparameters of the neural architecture.

Hyperparameter	Optimized	Fixed value	Value / value range
Batch size	X		$\{5, 10, 20, 25, 50\}$
BLSTM layer hidden neurons	X		$\{100, 101, ..., 255, 256\}$
BLSTM dropout rate	X		$\{0.2, 0.3, ..., 0.8, 0.9\}$
Learning rate		X	0.01
Max. number of training epochs		X	500
Patience for early stopping		X	20
Loss function		X	Negative Log Likelihood
Optimizer		X	AdamW

Table 3 – Composition of training, validation and test sets for the three different evaluations. All sets within the same evaluation are disjoint and stratified in relation to the target labels.

Evaluation	Training set		Validation set		Test set	
	Total words	Speakers	Total words	Speakers	Total words	Speakers
Single speaker	800	1	100	1	100	1
Leave-one-speaker-out	2400	3	600	3	1000	1
Multi-speaker	3200	4	400	4	400	4

For each type of evaluation, an ablation study was carried out by removing the lip and/or the lateral sensors. The neural networks were developed using the PyTorch library [15].

3 Results and discussion

The results of the three types of evaluations are shown in Tables 4 and 5. The following subsections will cover the command word recognition accuracies obtained using all sensors and how it compares to the state of the art in the literature, as well as the ablation study with lip and lateral sensors.

3.1 Command word recognition

The obtained single speaker and leave-one-speaker-out accuracies will be discussed using the study of Stone and Birkholz [6] as benchmark, which is the most direct comparison to this study, as the same neural architecture, the same number of speakers and a very similar recording device were applied.

The accuracies of the "best" test speaker obtained in this study are in a similar range than those from [6]: 93% and 59.70% from this study against 98.33% and 63.67%, for single speaker and leave-one-speaker-out evaluation, respectively. On the other hand, the mean accuracies (averaged across all 4 speakers) obtained in this study are in a somewhat lower range than those from [6]: 81.25% and 47.93% from this study against 97% and 56.17%, for single speaker and leave-one-speaker-out evaluation, respectively. The higher mean accuracies obtained by Stone and Birkholz [6] are higher than those obtained in this study probably due to two main reasons: the higher difficulty in using a corpus of 100 words and the poor performance of speaker 4 here in comparison to the other 3 speakers, which could be considered an outlier. After inspection of the data recorded by speaker 4, no visible measurements errors besides a high base value of one lip sensor (S12) were observed and the data was kept in the analysis.

The results here can also be compared to [7], where a single speaker command word recognition accuracy of 98.38% was achieved. This comparison is not as direct as the previous one: despite the exact same measuring device (OPG2023) was used, their corpus was smaller (40

Table 4 – Command word recognition accuracies obtained for single speaker and leave-one-speaker-out evaluations. The right portion of the table shows the mean and standard deviation (STD) values for each sensor set, as well as the absolute and relative differences in mean accuracy (Δ Ablation) compared to when all sensors are used.

Single speaker								
							Δ Ablation	
Sensor set	S1	S2	S3	S4	Mean	STD	Absolute	Relative
All sensors	89.00%	86.00%	93.00%	57.00%	81.25%	14.22%	-	-
No lip	90.00%	78.00%	90.00%	46.00%	76.00%	18.00%	−5.25%	−6.46%
No lateral	87.00%	82.00%	91.00%	9.00%	67.25%	33.78%	−14.00%	−17.23%
No lip/lateral	88.00%	81.00%	90.00%	50.00%	77.25%	16.08%	−4.00%	−4.92%
Leave-one-speaker-out								
							Δ Ablation	
Sensor set	S1	S2	S3	S4	Mean	STD	Absolute	Relative
All sensors	56.90%	53.80%	59.70%	21.30%	47.93%	15.51%	-	-
No lip	43.90%	40.40%	54.30%	20.60%	39.80%	12.21%	−8.13%	−16.96%
No lateral	51.40%	18.20%	16.00%	13.50%	24.78%	15.46%	−23.15%	−48.30%
No lip/lateral	45.10%	36.20%	55.10%	25.00%	40.35%	11.10%	−7.58%	−15.81%

Table 5 – Command word recognition accuracies obtained for the multi-speaker evaluation, and the absolute and relative differences in accuracy (Δ Ablation) obtained when different sensor set were used.

Multi-speaker				
	All sensors	No lip	No lateral	No lip/lateral
All speakers	90.25%	88.00%	88.50%	80.75%
Δ Ablation (absolute)	-	−2.25%	−1.75%	−9.50%
Δ Ablation (relative)	-	−2.49%	−1.94%	−10.53%

words) and a neural network was not applied.

Considering multi-speaker evaluation, our results stand in the same range as results obtained using different sensing modalities and neural architectures. The 90.25% accuracy from this study is most comparable with the 90.67% from Wu et al. [9], since the used corpora have the same size of 100 words. With smaller corpora, Zhang et al. [11] and Sun et al. [12] achieved 95.5% (31 words) and 93% (50 words), respectively.

3.2 Ablation study

The effect of removing the lip and/or the lateral sensors from the input data differed between the types of evaluation. In every evaluation, the removal of one or both sensor sets always resulted in a decrease in accuracy. The decreases in the multi-speaker evaluation were intuitively stronger when less articulatory information was available: removing both lip and lateral sensors resulted in a 9.50% accuracy decrease, whereas removing either lip or lateral sensors individually decreased accuracy by 2.25% and 1.75%, respectively.

However, in single speaker and leave-one-speaker-out evaluations, the stronger accuracy decreases occurred unexpectedly when only the lateral sensor was removed. This has two causes: the considerably worse performance achieved with data from Speaker 4, an outlier in comparison to the other 3 speakers, and the lower accuracies achieved when only lateral sensors were removed, in comparison to other sensor sets, for speakers 2, 3, and 4 (the first two only in the leave-one-speaker-out evaluation). This latter cause can be attributed to model underfitting, as none of the 30 hyperparameters sets resulted in accuracies if not equal, at least similar to the

case when neither lip nor lateral sensors were used.

In the single speaker evaluation, if only speakers 1, 2, and 3 are considered, the mean accuracies become 86.00%, 86.67% and 86.33% without lip, lateral and both lip and lateral sensors, respectively. These are more intuitive results, showing no sensible differences when different sensor regions are removed from the input data. In the leave-one-speaker-out evaluation, however, this trend persists even if Speaker 4 is not considered, as speakers 2 and 3 showed similar results.

To summarize, every sensor region investigated in the ablation study has a positive influence on the recognition accuracy, as the accuracy with all 15 sensors was in all cases the highest. Additionally, the presence of more speakers in the dataset seems to reduce the effect of the removal of individual sensors sets, either lip or lateral.

4 Conclusions

This paper presents the first command word recognition study using OPG2023 and data from multiple speakers. The obtained multi-speaker results are in the same range as other state-of-the-art SSI-based systems for command word recognition. This study, however, has considerably less speakers than most of these studies. Additionally, the multi-speaker accuracies were higher than those obtained in the single speaker evaluation, signaling potential for improvements if data from more speakers is added.

The ablation study with lip and lateral sensors suggested some conclusions. Despite there is a decrease in accuracy when lip and lateral sensors are not considered, the possibility of developing future versions of OPG2023 using less optical sensors could be explored. On the other hand, lip and lateral sensors seem to be important for generalization of recognition systems with multiple speakers. As future recordings with more speakers are a possible next step, these additional sensors should be considered.

5 Acknowledgements

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany in the programme of “Souverän. Digital. Vernetzt.”. Joint project 6G-life, project identification number: 16KISK001K.

References

[1] GONZALEZ-LOPEZ, J. A., A. GOMES-ALANIS, J. M. M. DOÑAS, J. L. PÉREZ-CÓRDOBA, and A. M. GOMEZ: *Silent speech interfaces for speech restoration: A review*. *IEEE Access*, 8, pp. 177995–178021, 2020. doi:10.1109/ACCESS.2020.3026579.

[2] REBERNIK, T., J. JACOBI, R. JONKERS, A. NOIRAY, and M. WIELING: *A review of data collection practices using electromagnetic articulography*. *Laboratory Phonology*, 12(1), 2021. doi:https://doi.org/10.5334/labphon.237.

[3] BIRKHOLZ, P. ET AL.: *A review of palatographic measurement devices developed at the TU Dresden from 2011 to 2022*. In R. SKARNITZL and J. VOLÍN (eds.), *Proceedings of the 20th International Congress of Phonetic Sciences*, pp. 883–887. 2023.

[4] FLETCHER, S. G., P. A. DAGENAIS, and P. CRITZ-CROSBY: *Teaching vowels to profoundly hearing-impaired speakers using glossometry*. *Journal of Speech, Language, and Hearing Research*, 3(4), 1991. doi:10.1044/jshr.3404.94.

[5] WAGNER, C., L. STAPPENBECK, H. WENZEL, P. STEINER, B. LEHNERT, and P. BIRKHOLZ: *Evaluation of a non-personalized optopalatographic device for prospective use in functional post-stroke dysphagia therapy*. *IEEE Transactions on Biomedical Engineering*, 69(1), 2022. doi:10.1109/TBME.2021.3094415.

[6] STONE, S. and P. BIRKHOLZ: *Cross-speaker silent-speech command word recognition using electro-optical stomatography*. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7849–7853. 2020. doi:10.1109/ICASSP40776.2020.9053447.

[7] POSSAMAI DE MENEZES, J. V., A.-L. FIETKAU, T. DIENER, S. KÜRBIS, and P. BIRKHOLZ: *A demonstrator for articulation-based command word recognition*. In *Proc. Interspeech 2024*, pp. 2042–2043. 2024.

[8] VORONTSOVA, D. ET AL.: *Silent EEG-speech recognition using convolutional and recurrent neural network with 85% accuracy of 9 words classification*. *Sensors*, 21(20), 2021. doi:10.3390/s21206744.

[9] WU, J., Y. ZHANG, L. XIE, Y. YAN, X. ZHANG, S. LIU, X. AN, E. YIN, and D. MING: *A novel silent speech recognition approach based on parallel inception convolutional neural network and mel frequency spectral coefficient*. *Frontiers in Neurorobotics*, 16, 2022. doi:10.3389/fnbot.2022.971446.

[10] WAGNER, C., P. SCHAFFER, P. AMINI DIGEHSARA, M. BÄRHOLD, D. PLETTEMEIER, and P. BIRKHOLZ: *Silent speech command word recognition using stepped frequency continuous wave radar*. *Scientific Reports*, 12(1), p. 4192, 2022. doi:10.1038/s41598-022-07842-9.

[11] ZHANG, R., K. LI, Y. HAO, Y. WANG, Z. LAI, F. GUIMBRETIÈRE, and C. ZHANG: *Echospeech: Continuous silent speech recognition on minimally-obtrusive eyewear powered by acoustic sensing*. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. Association for Computing Machinery, New York, NY, USA, 2023. doi:10.1145/3544548.3580801.

[12] SUN, X., J. XIONG, C. FENG, H. LI, Y. WU, D. FANG, and X. CHEN: *EarSSR: Silent speech recognition via earphones*. *IEEE Transactions on Mobile Computing*, 23(8), pp. 8493–8507, 2024. doi:10.1109/TMC.2024.3356719.

[13] WAGNER, C.: *Retainer-Free Optopalatographic Device Design and Evaluation as a Feedback Tool in Post-Stroke Speech and Swallowing Therapy*. Ph.D. thesis, Dresden, 2023.

[14] WELLS, J. C.: *Longman pronunciation dictionary*. Pearson Longman, Harlow, 3. ed., 3. impression edn., 2009.

[15] PASZKE, A. ET AL.: *Pytorch: An imperative style, high-performance deep learning library*. In *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates Inc., Red Hook, NY, USA, 2019.