Speaking Recognition with Facial EMG Sensors

Antonio Nikoloski[1], Petar Poposki[1], Ivana Kiprijanovska[2, *], Simon Stankoski[2], Martin Gjoreski[3], Charles Nduka[1], Hristijan Gjoreski[1, 2]

[1] Ss. Cyril and Methodius University in Skopje, N. Macedonia
[2] Emteq Ltd, Sussex Innovation Centre, Science Park Square, Brighton, UK
[3]Università della Svizzera Italiana, Switzerland
ivana.kiprijanovska@emteqlabs.com[*]

ABSTRACT

With the advent of interactive virtual reality (VR) applications, the interest in tools that allow users to engage with VR environments unobtrusively and intuitively is growing. One such interfacing tool for VR applications is speech recognition, which can contribute to enhanced human-computer interaction. In this study, we explore the usage of a novel VR facial mask equipped with seven surface electromyography (sEMG) sensors to recognize if the user is speaking or not using machine learning. We collected speaking and non-speaking data from 30 participants. The machine learning pipeline that was developed included data preprocessing, de-noising, filtering, segmentation, feature engineering, and training of a binary classification model. The experimental results indicate that the mask can be used to recognize the speaking activity. On the test data of five unseen participants, the best-performing model achieved an accuracy of 89% and an F1-macro score of 91. Additionally, by removing each sensor from the dataset, we analyzed the individual influence each sensor has on the models' outcomes. We did not observe a significant drop in the accuracy of the models, indicating that using the mask speaking can be detected even if some of the sensors are not used.

KEYWORDS

speaking recognition, machine learning, classification, wearable sensors, surface EMG, facial muscles.

1 INTRODUCTION

Virtual reality (VR) is an emerging technology that has introduced immersive user experience in virtual environments and is expected to revolutionize the way we interact with the digital world. VR applications have already been widely used in many different disciplines, ranging from research and training facilities to entertainment and healthcare. With the emergence of interactive VR applications, there is an increasing interest in new immersive tools that enable users to interact with VR surroundings in an unobtrusive and intuitive manner. One such interfacing tool for VR applications is speech recognition. Its incorporation with VR provides users with increased flexibility for interfacing with VR environments and can contribute to improved human-computer interaction.

In recent years, surface electromyography (sEMG)-based interfaces have been utilized for unobtrusive interaction in a VR environment. sEMG is used to measure muscle contractions using sensors applied directly on the skin by detecting changes in surface voltages on the skin when muscle activation occurs. In part due to its ability to be applied non-invasively, facial sEMG has been used to detect the activation of facial muscles that are activated during speaking. However, most sEMG sensors used in conventional speaking recognition systems have been attached around the user's lips and neck. This poses a number of practical issues, including the need for extra wearable devices in addition to the VR headset, limited facial muscle movement, and user discomfort.

To overcome these issues, in this study we explore the usage of a novel facial mask equipped with sEMG sensors. The mask is incorporated into a VR headset to recognize if the user is speaking or not. Our approach is based on signal processing and machine learning (ML), which are used to develop a binary classification model.

2 RELATED WORK

The first studies with sEMG sensors were performed by Piper[1]. Since then, researchers have been widely using sEMG sensors to measure the electrical signal that emanates from contracting muscles. The usefulness of the sEMG signal for measuring human performance was demonstrated by Inman [2] who investigated the technical aspects of human locomotion. By the early 1960s, the improvements in signal quality and convenience made the sEMG sensors a common tool in clinical and research laboratories. Despite their popularity, current recording methods can be problematic in maintaining signal fidelity when vigorous or long-duration activities are monitored [4] [3] .

Speech recognition by using sEMG was first used in the 80s [4] [6] . The results in these studies were preliminary but important for the further progress of the field. Jorgensen and Binsted [6] showed that it is possible to recognize speaking even if the words are spoken silently and/or without any actual sounds. Jou et al. [7] showed that it is possible to recognize not just the words but also the phonemes to a certain degree. Additional works include direct synthesis of speech via sEMG – which aids people who have problems with their vocal cords or airways [8] [9] .

Compared to the previous studies, we differ in the sense that we are using a novel facial mask – emteqPRO^{tm}, which is equipped with seven sEMG sensors. The sEMG sensors may be more error-prone compared to the intramuscular EMG sensors, and thus here we study their utility. Additionally, the location of our sEMG sensors makes the task of speaking recognition more challenging because the facial mask is placed on the upper part of the face (as part of the VR headset) and not the mouth and the lips – which would be more convenient for speech recognition.

3 DATASET

The data collection protocol included healthy participants that were asked to read a pre-defined text (news article). Additionally, we recorded a segment where the participants were sitting still, i.e., we recorded a baseline session with a neutral face. This data was recorded while the participants were watching a neutral video, without moving their facial muscles or speaking. A total of 30 participants were recorded, of which 18 were male and 12 were female, with a mean age between 19 and 25 years. The native language of all the participants was Macedonian.

During the data collection protocol, we were using the emteqPRO^{tm} mask [10] [11] to record sEMG sensor data. The mask has seven EMG sensors (Figure 1): two frontalis sensors (6 and 0 in Figure 1) used to monitor eyebrow movement; two orbicularis sensors (4 and 2 in Figure 1) used to monitor eye movements; two zygomaticus sensors (5 and 1 in Figure 1) used to monitor mouth and cheek movements; and one corrugator sensor (3 in Figure 1) used to monitor forehead movements.

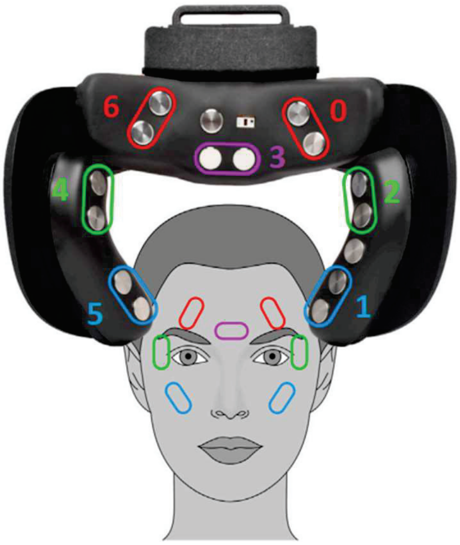

Figure 1: emteqPRO face mask with all 7 EMG sensors

4 DATA PREPROCESSING AND MODELING

The sEMG data were continuously recorded at a fixed rate of 1000 Hz. These data underwent a data preparation process, which included data filtering, segmentation, and feature engineering. To improve the quality of the sensor data, we performed signal de-noising and filtering. The EMG signals were initially filtered with a Hampel filter to eliminate sudden peaks in the signals that emerge as a result of quick movements. Additionally, we also applied a frequency-based filtering method based on spectrum interpolation [12] to reduce the noise caused by electromagnetic interference. [12] A sliding window technique was utilized for data segmentation. Specifically, the data were segmented into windows of size of 0.5 seconds with 0.4 seconds overlap (0.1 seconds slide). Finally, for each sEMG channel, we extracted 34 features, including various amplitude-based features, amplitude derivatives, auto-regressive coefficients, frequency-based features, and statistical features. The feature extraction procedure resulted in a total of 238 features.

The extracted features were used as input to four classification algorithms: (i) K- Nearest Neighbors [13] - a simple statistical algorithm where a datapoint is assigned a class according to the most numerous class of its k nearest neighbors; (ii) Support Vector Machine Classifier (SVM) [14] – an algorithm that works along the principle of finding a hyperplane in N-dimensional space to separate two classes of data points; (iii) Random Forest [15] - an ensemble learning method that trains N decision trees using random subsets of data and features and determines the instance's class by majority voting among the trained decision trees; and (iv) Extreme Gradient Boosting [16] - a gradient boosting algorithm which trains decision tree models sequentially, and each subsequent model strives to correct the errors of its predecessors.

5 EXPERIMENTS

5.1 Evaluation Setup

The recorded data was split into training (20 of the participants), validation (5 of the participants) and test datasets (5 of the participants). The train dataset was used to train the models, the validation was used to optimize hyperparameters, and the test dataset was used to report the accuracy. The evaluation metrics we used to test the performance of our models were accuracy and F1 score.

Additionally, the experiments were performed so that the training validation and test subsets do not have overlapping participants - i.e., each participant's data is found only in one of the three subsets. This is done so that we replicate a scenario where the model is used in practice on participants that are not in the training dataset.

5.2 Default Hyperparameters Results

Figure 2 presents the results (accuracy and F1-score) achieved by each of the algorithms with their default hyperparameters. We additionally included the Dummy classifier as a reference (which predicts the majority class). The results show significant improvement by all the algorithms compared to the Dummy classifier. The Random Forest and the SVM achieved similar results, while the XGBoost classifier achieved the best results overall (87% accuracy and 89% F1-score). Apart from this, this classifier also scaled efficiently with the size of the datasets, as it was able to quickly and efficiently create and train models. This was also beneficial for the hyperparameter optimization – explained in the next subsection.

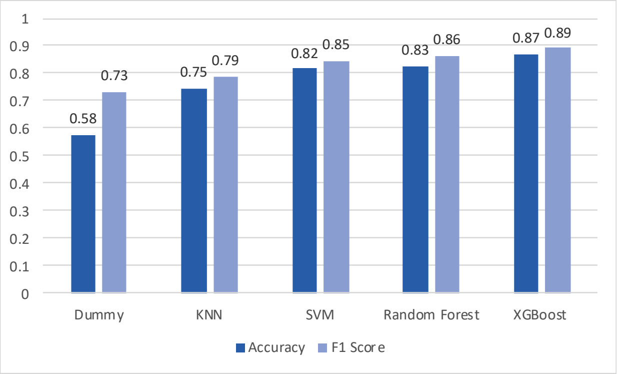

Figure 2: Algorithm comparison (accuracy and F1-score) using default hyperparameters

5.3 Optimized Hyperparameters Results

In the next step, we performed hyperparameter optimization. This process involves iterative changes of certain parameters of a classifier. During this process, an interval for every hyperparameter is defined, and afterward, each parameter is iteratively updated, and the performance of the models is monitored. During this step, all 238 features of the datasets were used, and a large number of numerical and other parameters (such as kernel for SVM, booster for XGB, etc.) were tuned.

Figure 3 presents the results (accuracy and F1-score) achieved by each of the algorithms after the hyperparameter optimization. The results show slight improvement for the KNN, SVM, and XGBoost algorithms, the latest one achieving 89% accuracy and 91% F1-score – which was the best score that we achieved on this dataset.

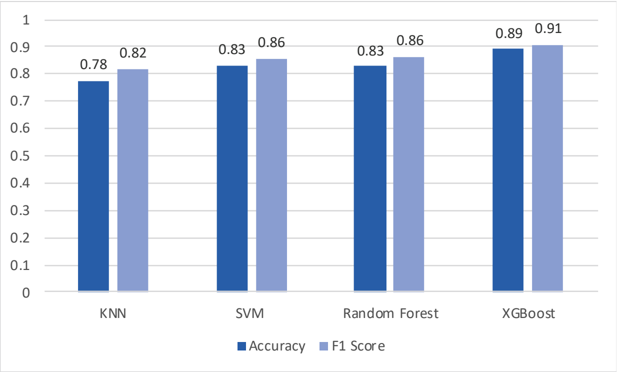

Figure 3: Algorithm comparison (accuracy and F1-score) using optimized hyperparameters

5.4 Continuous Recognition Results

Figure 4 illustrates the continuous recognition results for the five subjects from the test set achieved by the best-performing XGBoost classifier. A comparison was made between the true and the predicted class on a time scale, i.e., with a blue line, the true classes are presented (1 represents speaking, 0 represents not speaking). Additionally, the orange color presents the speaking predictions by the model. Each subject's data is separated with black dashed lines in the figure. The results show that a large portion of the error is down to the baseline sessions of the last two subjects in the test dataset, marked with red circles. In a large portion of the baseline sessions, the model is falsely predicting speaking activity. We speculate that the reason might be that these two subjects were moving their head during the baseline session, which may have caused the sensors to shift from their original position and deteriorate their contact with the skin.

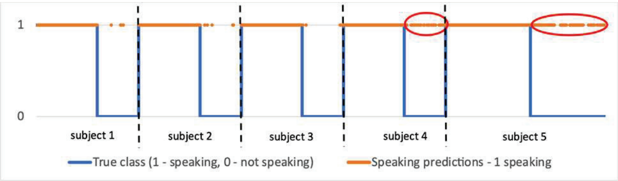

Figure 4: Continuous recognition results for the XGBoost algorithm. The blue line represents true classes (1 – speaking, 0 – not speaking), and the orange line represents the predictions (1 – speaking)

5.5 Sensor Analysis Results

We additionally analyzed the results achieved by the models if a certain sensor is missing. This way, we were able to check the importance of each sensor for the given task. Knowing the positions of the sensors on the face, we wanted to learn how the data would change if we were to drop data from a certain sensor while keeping the rest.

The results are shown in Figure 5, which in general, show that the drop in accuracy and F1 score is not significant for all the sensors. The accuracy drops from 87% to 85% at most. A more detailed analysis shows that the sensors placed on left and right orbicularis, corrugator, and left frontalis have the most impact on accuracy, i.e., the accuracy drops the most when one of these sensors is missing. One of the reasons for this is that while the participants were speaking, they were actually reading – which means they activated their eyes which is recorded by the orbicularis muscles. This analysis shows us that certain muscles activate more while speaking compared to others, so that is why the model itself gains or loses accuracy more, depending on which sensor is dropped.

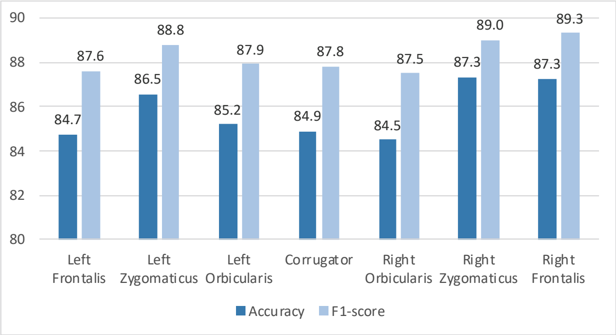

Figure 5: Sensor analysis showing the performance when a particular sensor is missing.

6 CONCLUSION

In this work, we presented a ML approach for speaking recognition using facial sEMG sensors integrated into a VR headset. The dataset was collected with 30 healthy participants while reading a news article and watching videos. The results show that the best performing model is XGBoost, which achieved 89% accuracy. Additionally, the error analysis per participant showed that most of the misclassifications were incorrect speaking predictions in the baseline (non-speaking) sessions of two participants. We speculate that this is caused by the head movement of the participants and we plan to tackle this using the IMU sensor on the emteqPRO^{tm} mask.

An additional problem was that while the participants were reading, they were making small breaks, which were automatically labeled as speaking – but in fact were not speaking. This labeling problem will be tackled in future by using audio to exactly label the speaking segments.

Finally, we plan to implement person-specific normalization on the EMG data. This is an important step given that different participants have different facial muscles, and even more, those muscles are activated differently while doing the same facial expressions or speaking.

ACKWNOLEDGEMENT

Part of this study was supported by the Innovate UK Project no. 81376: Virtual Reality rehabilitation tailored to older brain injury patients (Healthy Ageing), and part by the WideHealth project no. 952279 - European Union's Horizon 2020 research and innovation programme.

REFERENCES

[1] Piper H (1912) Elektrophysiologie menschlicher Muskeln. Springer, Berlin, pp 1–163.

[2] Inman, V. T., Saunders, J. B., & Abbot, L. C. (1944). Observations on the function of the shoulder joint. Journal of Bone and Joint Surgery, 26, 1-30.

[3] M. Wand, M. Janke, and T. Schultz, "Investigations on Speaking Mode Discrepancies in EMG-based Speech Recognition," in Proc. Interspeech, 2011, pp. 601–604.

[4] N. Sugie and K. Tsunoda, "A speech prosthesis employing a speech synthesizer—Vowel discrimination from perioral muscle activities and vowel production," IEEE Trans. Biomed. Eng., vol. BME-32, no. 7, pp. 485–490, Jul. 1985.

[5] M. S. Morse and E. 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," Comput. Biol. Med., vol. 16, no. 6, pp. 399–410, 1986.

[6] C. Jorgensen and K. Binsted, "Web browser control using EMG based sub vocal speech recognition," in Proc. 38th Annu. Hawaii Int. Conf. Syst. Sci., 2005, p. 294c.

[7] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in Proc. Interspeech, 2006, pp. 573–576.

[8] J. Freitas, A. Teixeira, and M. S. Dias, "Towards a silent speech interface for portuguese," in Proc. Biosignals, 2012, pp. 91–100. [23] A. Toth, M. Wand, and T. Schultz, "Synthesizing speech from electromyography using voice transformation techniques," in Proc. Interspeech, 2009, pp. 652–655.

[9] K.-S. Lee, "Prediction of acoustic feature parameters using myoelectric signals," IEEE Trans. Biomed. Eng., vol. 57, no. 7, pp. 1587–1595, Jul. 2010.

[10] Gjoreski, H., I. Mavridou, I., Fatoorechi, M., Kiprijanovska, I., Gjoreski, M., Cox, G., & Nduka, C. EmteqPRO: Face-mounted Mask for Emotion Recognition and Affective Computing. In Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (pp. 23-25).

[11] Gnacek, Michal & Broulidakis, John & Mavridou, Ifigeneia & Fatoorechi, Mohsen & Seiss, Ellen & Kostoulas, Theodoros & Balaguer-Ballester, Emili & Kiprijanovska, Ivana & Rosten, Claire & Nduka, Charles. 2022. emteqPro-Fully Integrated Biometric Sensing Array for Non-Invasive Biomedical Research in Virtual Reality. Frontiers in Virtual Reality. 3. (Mar. 2022)

[12] Mewett, D. T., Reynolds, K. J., & Nazeran, H. Reducing power line interference in digitised electromyogram recordings by spectrum interpolation. Medical and Biological Engineering and Computing, 42(4), 524-531, (2004).

[13] D. Aha, D. Kibler (1991). Instance-based learning algorithms. Machine Learning. 6:37-66.

[14] Zhang, Yongli. (2012). Support Vector Machine Classification Algorithm and Its Application. 179-186.

[15] Breiman, "Random Forests", Machine Learning, 45(1), 5-32, 2001.

[16] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794.