applied sciences

MDPI

Article

A Novel Approach to EEG Speech Activity Detection with Visual Stimuli and Mobile BCI

Marianna Koctúrová and Jozef Juhár *

Department of Electronics and Multimedia Communications, Technical University of Košice, 042 00 Košice, Slovakia; marianna.kocturova@tuke.sk
* Correspondence: jozef.juhar@tuke.sk

Abstract: With the ever-progressing development in the field of computational and analytical science the last decade has seen a big improvement in the accuracy of electroencephalography (EEG) technology. Studies try to examine possibilities to use high dimensional EEG data as a source for Brain to Computer Interface. Applications of EEG Brain to computer interface vary from emotion recognition, simple computer/device control, speech recognition up to Intelligent Prosthesis. Our research presented in this paper was focused on the study of the problematic speech activity detection using EEG data. The novel approach used in this research involved the use visual stimuli, such as reading and colour naming, and signals of speech activity detectable by EEG technology. Our proposed solution is based on a shallow Feed-Forward Artificial Neural Network with only 100 hidden neurons. Standard features such as signal energy, standard deviation, RMS, skewness, kurtosis were calculated from the original signal from 16 EEG electrodes. The novel approach in the field of Brain to computer interface applications was utilised to calculated additional set of features from the minimum phase signal. Our experimental results demonstrated F1 score of 86.80% and 83.69% speech detection accuracy based on the analysis of EEG signal from single subject and cross-subject models respectively. The importance of these results lies in the novel utilisation of the mobile device to record the nerve signals which can serve as the stepping stone for the transfer of Brain to computer interface technology from technology from a controlled environment to the real-life conditions.

Keywords: electroencephalography; mobile EEG device; speech detection; feed-forward neural network; visual stimuli

check for updates

Citation: Koctúrová, M.; Juhár, J. A Novel Approach to EEG Speech Activity Detection with Visual Stimuli and Mobile BCI. *Appl. Sci.* **2021**, *11*, 674. https://doi.org/10.3390/app11020674

Received: 30 November 2020
Accepted: 5 January 2021
Published: 12 January 2021

1. Introduction

The interest in the field of speech detection and recognition has been on the increase through the last decade thanks to the new possibilities of its application in the technology that can improve our lives. New speech recognition approaches focus on communication without an audio signal. Technologies which use other signals to detect speech processes include electromyography for recording the movement of facial and articulatory muscles, invasive electrocorticography for recording electrical activity on the surface of the brain, or electroencephalography for recording brain electrical activity from the surface of the head, and others. These technologies represent a wide range of possibilities for speech recognition and detection for the physically or mentally disadvantaged and for their use in environments where sound cannot be used. Possibilities of using speech detection on the revelation of a speech pattern of the brain are based. That pattern can be excluded when examining other brain activities or used to improve speech recognition from brain activity. Following international publications and studies concerning electroencephalography (EEG) technology served as the background research for our own initiative.

In Reference [1], the non-speech state of the brain activity was examined. The research focused on the existence of silence signatures in EEG and its identification. The main area of examination is the classification of brain activity regions during heard and imagined speech. The best accuracy of the classifier detection was about 50%. In the work, the 4-channel Muse EEG headset and a 128 channel EGI device were used.

Automatic real-time voice activity detection was examined in Reference [2]. This study was focused on magnetoencephalography (MEG) signals. Data was collected by the NeuroVAS device in a medical environment. The study offers a method of classifying speech and silence intervals using a support vector machine, which achieves an average accuracy of 88%.

Further studies investigating speech detection from MEG signals [3] used the Gaussian mixture model and artificial neural network. In the experiment, the speech was detected on whole phrases. The best classification accuracy was achieved at 94.54%.

From a neurological point of view, we can categorise human speech according to the role it represents. The similarities in EEG signals in the induction of various speech activities such as perception, production, and imagination of speech is the subject of an experimental study [4]. Studies reported results higher than the chance level for EEG classification using machine learning with a data set that includes 30 subjects.

One of the first online classifications to covert speech from EEG signals has been described in Reference [5]. This work deals with the distinction of the mentally spoken words. The research involved 20 subjects who were tasked to mentally repeat the words “yes” and “no”. Using support vector machine (SVM) classification an average accuracy of 75.9% was achieved.

The use of speech activity detection from EEG is discussed in Reference [6]. The experiment deals with the improvement of speech recognition in a noisy environment using EEG signals. The data sets used for the experiment were recorded in a noisy environment, which was classified using Recurrent Neural Network. The results showed that the EEG signal can help improve the voice activity detection and can also be helpful as a predictor of the decision to talk or not.

In our research, we dealt with the detection of speech that is associated with visual stimuli. The classification of the state of inactivity and the state when the subject was exposed to a visual stimulus has been described in Reference [7]. This paper details an experiment where the researched subjects were presented in a visual stimuli in the form of a single colour image on the screen followed by a state of relaxation. The output EEG brain signals were measured and classified to detect a state of inactivity and an exposure to a visual stimulus state. The classifiers used were SVM and random forest. The result of the classification of the state of inactivity and the state of exposure to the image stimulus was on average 94.6%.

A precursor work to this research was carried out in 2017 [8]. This work was aimed at the recognition of the commands from the EEG signal, however only low level of accuracy was achieved. The data set of EEG signals was compiled using 20 subjects who pronounced 50 different commands and a cross-subject Hidden Markov model was trained to recognize the spoken commands. The best recognition accuracy was more than 5%, which is more than the chance level, but many false alarms occurred during the recognition. However, this work had paved the way to the idea of improving the speech recognition from EEG signals by adding a speech detector that can increase the overall accuracy.

The above work was followed by a research summarised in Reference [9], This research was focused on the speech activity detection from EEG using the Feed-Forward Neural network with a single-subject model. The research has achieved a considerable improvement in the detection accuracy averaging at 77% for F1 results.

2. Motivation

In this work is aimed at the investigation of the speech activity present in the EEG signals. The main focusing on the cooperation of several areas of the brain, especially speech activity that is associated with visual stimuli. The stimuli used in the experiment were images and written text. According to the study in Reference [10], the visual stimulus can generate inner speech. In speech communication, a person uses the visual area of the brain which accelerates the speech generation. The human brain often works with the visual or textual form of objects, to identify the correct grammatical, audio, and articulatory form of the word. Such cooperation of several areas of the brain improves the communication process.

As mentioned we expected increased activity of the two main areas of the brain during EEG signal acquisition which are the speech area, and the visual area. During EEG recording the subjects were presented with both images with showing specific colours and the text forms of individual colours. By presenting the subjects with these images their visual and imaginary parts of the brain became activated. The important fact to realised is that the brain areas which control what we see and what we can imagine are the same. When communicating, we often involve these visual areas in addition to speech areas. The relationship between imagination and vision is very close and hence these two actions elicit similar signals. The main difference between visual and mental images lies in the communication of the cerebral pathway, which started from the eye, leads to the primary visual cortex [11,12]. The study in Reference [13] assumes that image recognition and image naming involves the activity of occipitotemporal and prefrontal areas.

Speech production is created by a so-called conversational loop that moves the information signal to different parts of the brain tin order to compose a word pattern before it is pronounced. Such a phonological loop involves the processing of the linguistic-auditory signal which is in the first instance formed in the auditory cortex followed by the transfer of the audio form of the word to the Wernicke's area. In the Wernicke's area the sound pattern is processed and from there it travels to the Broca's area where syntax, morphology, and articulation instructions are generated. This information is sent to the motor cortex finally where the movements of the vocal cords, tongue, diaphragm, and other voicing muscles are controlled [14].

3. An Experimental Database

3.1. Materials

Although considerable research has been done in modeling human speech detection and recognition, it is difficult to use this information to create a speech recognition model that works on a mobile EEG device. Working with mobile devices is often more difficult due to the inferior signal acquisition conditions. However, mobile EEG devices bring many benefits to end-users such as lower price, easier connection, easier to place on the head or the use of dry electrodes, to name few. In our research the database of EEG signals was recorded using the OpenBci Ultracortex Mark III EEG headset. Bluetooth signal was used to both control the mobile EEG head cap via a computer and to transfer the recorded EEG signal. The sampling frequency of 125 Hz was used to capture the brain signal via the EEG headset.

The headset configuration consisted of 16 dry electrodes that recorded the signals in different areas of the brain. The channels used were namely Fp1, Fp2, F3, F4, F7, F8, C3, C4, T3, T4, P3, P4, T5, T6, O1, O2, according to the international 10/20 system. The decision to apply the electrode configuration to the entire head area was made based on the reasons given in Section 2. Since the research focused on the brain activity during an overt speech associated with visual stimuli it is important to capture the signals coming from areas of the brain not just the speech area. Figure 1b shows the side view of the location of the sensors use in our research and the iodine of the brain is shown. Due to the limited possibilities of positioning the electrodes on the Ultracortex EEG headset, we tried to cover the following areas: Speech areas (Broca's and Wernicke's area), Auditory areas, Visual areas and Motor areas. Another reason to use EEG signals coming from multiple areas is that, despite the well-known knowledge that Broca's and Wernicke's areas are located on the left hemisphere in about 95% of right-handers and 70% of left-handers [15], observations in the previous experiment [8] did not show a significant improvement for the machine learning based on the modeling with left hemisphere EEG signals. The experiment in Reference [14], showed that the classification of EEG signals from the left hemisphere shows just slightly better accuracies compared to signals from the right hemisphere signals classification.

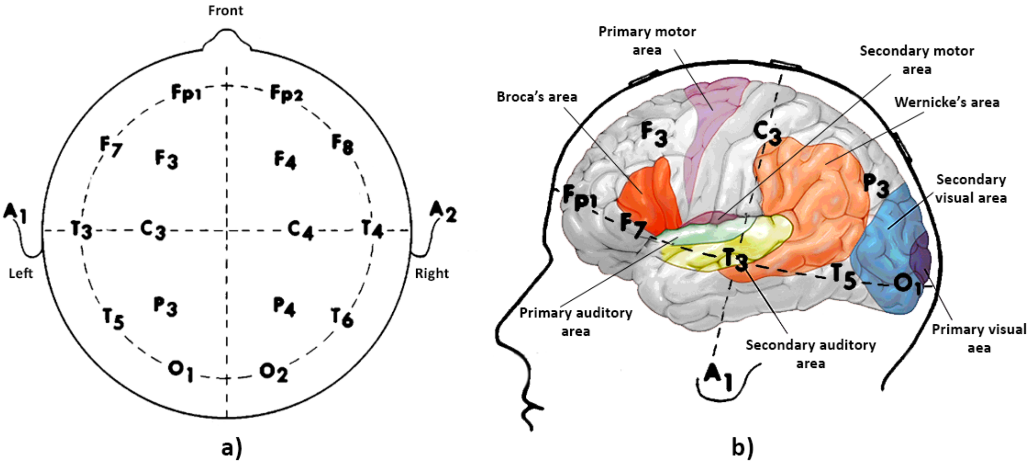

Figure 1. (**a**) Electrode configuration according to 10/20 international system used in the experiment. (**b**) Lateral view of the electrode configuration, compared with the brain areas (motor, auditory, visual, and speech).

Our EEG headset has been modified to use soft electrodes to improve the subject's comfort. Fitted measuring electrodes are based on a flexible conductive, elastic main body with a conductive coating covering the contact area. The functionality of the electrodes with the OpenBCI headset was discussed in Reference [16]. Our solution for the connection of the electrodes to the headset resulted in the snap connectors being soldered to wires to which electrodes were connected.

The EEG signals was recorded on four healthy, right-handed native Slovak speaking subjects. They signed an informed consent form in which they were acquainted with the purpose of the experiment and with the use and management of the personal data. The research was carried out in accordance with the Code of Ethics for employees of the Technical University in Košice.

During the experiment the subjects followed an experimental protocol. Their task was to sit motionlessly in a comfortable position while focusing on the screen in front of them and processing the word pronunciation. During the EEG signal recording the audio signal was also gathered, which were used for speech labels creating.

3.2. Acquisition Protocol

An acquisition protocol for EEG signals recording was partly inspired by the experiment in Reference [7]. The acquisition protocol was designed to activate the speech, visual and imaginary areas of the brain. EEG signals recording was performed in two sessions, each lasting about 10 min. In the first session specific colours were displayed on the screen as pictures, while in the second session, the colours were displayed as text on the screen. The recorded subject was tasked to focus on the displayed colours and to name them. In the first session the focus was on the activation of the subject's imaginary areas of the brain. These areas help to create the image form of the word or to transform the seen image into a verbal form. The second session of the experiments was focused on activating the area of the brain responsible for understanding the written text, where the written text is transformed into a verbal form and vice versa. In the above exercise, 10 easily distinguishable colours were displayed randomly one by one with short breaks represented by a plain black colour. The colours spectrum selected for this protocol were yellow, green, red, blue, orange, violet, white, pink, grey, and brown.

4. Methods

The audio and EEG signals were recorded together. The audio recording was used to create speech and non-speech labels. EEG signals recorded using the graphical user interface OpenBCI, and stored in a text file format, where the individual channels are written in columns. The file format also contains an additional column which contains the information about the Unix time of recording each sample. The main challenge was in the synchronisation of the recorded EEG and audio signals in the post processing. To solve is was created a simple script which, after starting the audio recording, automatically stored the Unix time of record start. This was used to synchronize an audio with EEG in post-processing.

For the experiment, 9 basic features was assembled based on which the models in the shallow Feed-Forward Artificial Neural Network was created. The important part of the signal processing was the conversion of the signal into minimum-phase signal. Ttures were calculated for the raw EEG signal as well as for the minimum-phase signal. In summary, for the total of 16 recorded EEG channels in the experiment a calculation was completed for ehe assigned feaach channel identifying 9 features from the raw signal and 9 features for the minimum-phase signal. All input features composed an input vector of size 288.

4.1. Labeling

The voice of each subject was captured during the EEG recording. The audio recording was used to identify and time aligns labels. The audio recordings were downsampled to match the sampling frequency of the EEG record and the individual speech and non-speech state samples were marked as 1 or 0.

4.2. Signal Pre-Processing

The recording of EEG signals according to the protocol described in Section 3.2 was performed. The pronunciation of individual words was alternated by longer pauses of inactivity. Parts with pronunciation are considered to be an active state of the brain when speech is produced, and pauses between words are considered to be the non-speech state. The average pronunciation time for one word was 0.75 s, while the average pause time lasted up to 4.25 s (Figure 2) and the average number of words in one recording was 120. In summary, it is clear that the speech time in one recording was about 90 s and the silence time was about 510 s. As a result, the overall length of the silence segments dominates significantly. Such a disproportion in the data would causes the model to be trained on the dominant target (non-speech) which would result in the creation of an erroneous and unacceptable speech detection model. To overcome this issue the data were balanced by random undersampling of majority labels (non-speech state labels) before further processing. Indexes of samples belonging to the zero labels were selected based on the indexes of the selected zero samples, the samples from both the EEG and the label data were marked and deleted, thus shortening the total length of the signal, but the data were balanced.

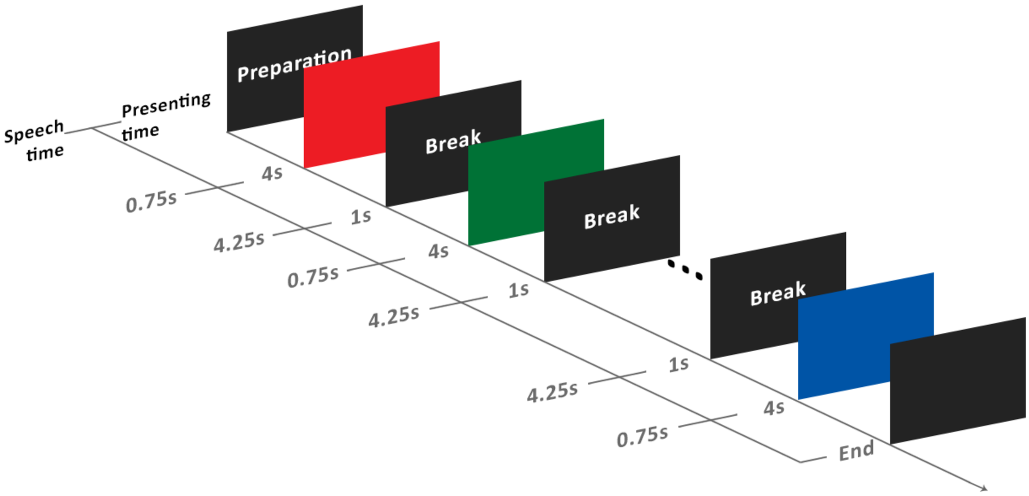

Figure 2. Protocol for electroencephalography (EEG) signal acquisition. On-screen presentation times and average speech times. Individual colours or colour names were displayed with breaks created by black colour on screen.

4.3. Minimum-Phase Signal Calculation

The first step before calculating the features that would define the signal was to calculate the minimum-phase signal from the obtained EEG recordings. Creating this equivalent signal is one of the features we want to use to improve speech detection. From a theoretical point of view, we can describe the minimum-phase signal as follows.

System with a minimal phase has the poles and zeros of its rational transfer function in the domain of the transformation $\mathbb{Z}$ lie inside a unit circle in a complex plane. The minimum-phase signal is defined as a signal in which the energy is concentrated near the front of the signal [17].

The method for obtaining the mminimum-phase phase equivalent uses the conversion of a signal to its real cepstrum using fast Fourier transforms and a recursive process from its real cepstrum backwards. The real cepstrum is the inverse Fourier transform of the real logarithm of the magnitude of the Fourier transform of a sequence [18]. First of all, we calculated the real cepstrum from the signal $x(n)$. The real cepstrum analysis can be expressed as [19]:

$$c_r(n) = \Re(\mathcal{F}^{-1}[\log|\mathcal{F}\{x(n)\}|]). \tag{1}$$

By a recursive procedure can be found the minimum-phase equivalent of a signal $x(n)$ from its real cepstrum $c_r(n)$:

$$x_{min}(n) = \Re(\mathcal{F}^{-1}\exp[\mathcal{F}\{c_r(n)\}]). \tag{2}$$

From the obtained minimum-phase signal $x_{min}(n)$, the features were calculated according to the following chapter.

4.4. Feature Extraction

Selected features were calculated to a signal for a window size of 10 signal samples with 50% overlap. The features were calculated for a pre-prepared EEG signal, as well as for the minimum-phase signal. The individual features extracted from the signal $x(n)$ are listed in the next lines [20].

- *Mean* value of the frame is an average value of signal parties

$$\mu = \frac{1}{N} \sum_{n=1}^{N} x(n). \tag{3}$$

- *Standard deviation*

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{n=1}^{N} (x(n) - \mu)^2}. \tag{4}$$

- *Skewness* indicates the measured the asymmetry of this distribution around its mean value. The skewness is defined for a signal $x(n)$ as:

$$Skew = \frac{1}{N\sigma^3} \sum_{n=1}^{N} (x(n) - \mu)^3, \tag{5}$$

  where μ and σ are the mean and standard deviation of the signal $x(n)$. Distribution of EEG signal is symmetric if it looks equal on the right and left sides of the centre [21].
- *Kurtosis* is the fourth-order central moment of distribution. EEG signals coefficients have not normal distribution and have heavy-tailed characteristic justified by kurtosis [21]. The kurtosis is defined for a signal $x(n)$ as:

$$Kurt = \frac{1}{N\sigma^4} \sum_{n=1}^{N} (x(n) - \mu)^4. \tag{6}$$

  Kurtosis characterises comparison of relative peakness distribution and normal distribution. Positive kurtosis coefficients indicate peaked distribution and the low or negative coefficients mean a flat envelope [22].
- *Energy* indicate the strength of the EEG signal. Energy used for our signal was calculated as the sum of all the squared values in the window. The average energy is define by:

$$Energy = \frac{1}{N} \sum_{n=1}^{N} x(n)^2. \tag{7}$$

- *Bandpower* is a method of monitoring amplitude modulations at the frequency range [23].

$$P = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)^2 dt. \tag{8}$$

- *Root mean square*

$$RMS = \sqrt{\frac{1}{N} \sum_{n=1}^{N} x(n)^2}. \tag{9}$$

- *Shannon entropy* of a signal is a measure of the spectral distribution of the signal. The Shannon entropy is measure of uncertainty in a random process [24]. It is defined as:

$$Shannon = \sum_{n=1}^{N} x(n) log_2 x(n). \tag{10}$$

- *Spectral flux* describes sudden changes in the frequency energy distribution.It indicates the rate of change of the power spectrum. Spectral flux is given by formula:

$$Spectralflux = \sum_{k=1}^{M/2} (S_t(k) - S_{t-1}(k))^2, \tag{11}$$

where $S_t(k)$ and $S_{t-1}(k)$ are the Fourier transforms of actual frame t and pervous frame $t-1$ of the signal [25,26].

4.5. Training Algorithm

We assume that machine learning models such as feed-forward neural network may identify patterns in the preprocessed EEG data to predict the desired speech and non-speech classes. The preprocessed data fed into the feed-forward neural network are transformed by a number of weights which form the output prediction for each class. At the beginning of the model training phase, all model weights are randomly initialized which results in the erroneous output. Later in this phase, the manually labeled classes, which are paired with the input data, are fed to the feed-forward neural network model. The error backpropagation algorithm is used to optimize the model weights in order to reduce the output error.

In our experiment, the 2-layer feed-forward neural network was used. Min-max scaling was performed by network pre-processing. The neural network consists of a single tanh activated hidden and binary output sigmoid activated output layer. We trained the network with Scaled conjugate gradient backpropagation, with binary cross-entropy as the loss function. Output pseudo probabilities were thresholded with a 0.5 decision boundary. The first layer of the network consisted of one hundred hidden neurons.

The models were created based on 18 features calculated for the frame of 10 signal samples. In previous experiments, it turned out that such a window size is sufficient and does not unnecessarily reduce the amount of information in the signal. In the experiment, a model was created using machine learning modelling, which would be able to recognize the state of speech production and the state of brain inactivity. Therefore, in the above classification, all words were explicitly considered as one class.

The experiment was divided into two main parts, in the first we tried to create single-subject models. In the second part, we created cross-subject models.

5. Experimental Results

In this section, the results of neural network testing are presented. As discussed in the main body of this report the experiment into was divided into two main part. In the first part the speech detection models for individual subject data set were created. The second part. In the second part the focus was on the creation of a model from the cross-subject data set. One hundred models for each data set were trained using the assembled neural network. From all of the models created the one with the highest result in the testing process was selected and the results with the highest values for data set models are presented here.

5.1. Model Evaluation Metrics

Scoring performance metrics were calculated to evaluate the achieved results. The data set may be slightly unbalanced due to signal processing where we introduce framing for calculating features the number of ones and zero targets may not agree. Therefore, the accuracy metric cannot be the only measure used to evaluate the results obtained. What we were aiming for was to find positive results for the speech activity state, so we consider the F1 score to be the most important measure and for completeness, we also provide recall and precision metrics.

1. *Recall*: is the ratio of true positive predictions to all positive predictions. Recall is also known as the True Positive Rate or the sensitivity. We can express it by:

$$Recall = \frac{TP}{TP + FN}. \tag{12}$$

2. *Precision*: is the actual correct proportion of positive predictions. This can be expressed by:

$$Precision = \frac{TP}{TP + FP}. \tag{13}$$

3. *F1 score*: is the harmonic mean of the precision and recall. We can express it by:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{14}$$

5.2. Single-Subject Model Experiment

In the first part of the experiment, models for the single-subject models were created, 70% of the data was used for training, 15% for the model validation, and 15% for the test.

As can be seen in Table 1, machine learning learned to detect speech from the EEG signal for a single subject with an average accuracy of 84%. This accuracy can be considered relatively high.

Table 1. Experimental results for test set single-subject model.

Subject Number	**Precision [%]**	**Recall [%]**	**Accuracy [%]**	**F1 Score [%]**
1	87.61	89.11	86.65	88.36
2	73.76	85.16	74.72	79.05
3	89.84	90.68	88.75	90.26
4	88.64	90.43	88.35	89.53
Average	84.96	88.85	84.62	86.80

The best result was recorded for the subject 3. Figure 3 shows the part of the input labels compared to the output data obtained from model based on data set 3.

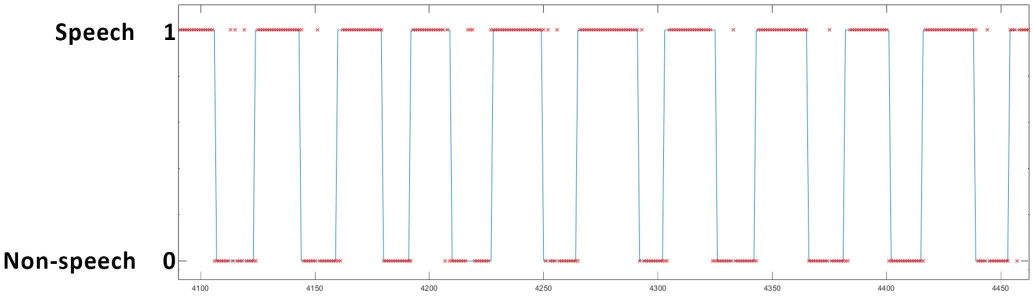

Figure 3. Graphical representation of speech labels (a blue shape) compared with predicted labels (a red crosses) within the range 0 to 1.

5.3. Cross-Subject Model Experiment

In the second part, the aim was to create a cross-subject and a cross-session models. Creating a model using the EEG signals from the multiple subjects is often very challenging. Revealing brain activity that shows the same signals for several people encounters a problem caused by subjective perception. Nevertheless, it can be assumed that if human speech is created and controlled in the same areas of the brain, it may also show similar brain waves, and these similarities can be revealed by machine learning.

In this part we selected data sets from combination of 3 subjects which were divided into 80% training set and 20% validation set. The created model was tested on the data set from the 4th subject.

5.3.1. Colour Naming Session

Table 2 lists the machine learning results for cross-subject models. In this experiment, we created a model for speech detection using signals from 3 subjects and tested on the rest quarter. The data used for these models were selected only from the first session, where the subjects were presented with colours in the form of an image and they named them.

Table 2. Cross-subject model, colour naming session.

Subject Numbers Model Training	Subject Number Model Testing	Precision [%]	Recall [%]	Accuracy [%]	F1 Score [%]
1, 2, 3	4	65.71	92.03	69.44	76.68
2, 3, 4	1	67.78	88.09	69.97	76.61
1, 3, 4	2	58.55	96.31	59.40	72.82
1, 2, 4	3	71.68	88.11	74.47	79.05
Average		65.93	91.14	68.32	76.29

5.3.2. Reading Session

Table 3 shows a results for cross-subject models created of data sets from the second session, where subjects were presented with the names of pictures in textual form and their task was reading.

Table 3. Cross-subject model, reading session.

Subject Numbers Model Training	Subject Number Model Testing	Precision [%]	Recall [%]	Accuracy [%]	F1 Score [%]
1, 2, 3	4	80.50	87.14	81.09	83.69
2, 3, 4	1	58.32	98.94	59.69	73.38
1, 3, 4	2	69.66	54.76	62.02	61.32
1, 2, 4	3	66.65	93.30	70.56	77.65
Average		68.78	83.54	68.34	74.01

5.3.3. Cross-Session Experiment

Table 4 was compiled based on results for cross-session speech detection. The data used for creating cross-session detection model were mixed from all subjects and both sessions.

Table 4. Cross-session model.

Subject Numbers Model Training	Subject Number Model Testing	Precision [%]	Recall [%]	Accuracy [%]	F1 Score [%]
1, 2, 3	4	69.61	92.67	73.71	79.50
2, 3, 4	1	60.14	95.73	62.14	73.87
1, 3, 4	2	61.70	62.81	57.48	62.25
1, 2, 4	3	70.37	86.37	72.56	77.56
Average		65.46	84.40	66.47	73.30

6. Discussion and Conclusions

Our research work has demonstrated the possibility of the speech detection using a novel mobile EEG device which can be utilised in the real-life conditions. Up to now, there has been a widely accepted scepticism regarding the potential accuracy of the speech detection based on the EEG signal used outside the controlled environment. The main argument against the use of such devices in the real-life conditions was based around the potential problems associated with the insufficient continuous connection for all channels, lower sample rate or difficult data transmission via wireless protocol. Our research has demonstrated good accuracy of such technology based our experiment which was focused on the speech detection through EEG signals using Feed-Forward Neural Network.

Our experimental work was divided into two main parts. The first part of the experiment was focused on the generation of the Neural Network models for individual subjects. The detection results for our EEG database were very positive for our research with an average F1 score of 86.8%. Based on the above results, we can see an improvement in the speech detection system in comparison with the previous experiment [9], which achieved F1 score of 77% for the single-model.

In the second part the data sets were set up for cross-subject speech detection. The cross-subject detection was considered to be an important step in our research. In this case, we expected that the combination of the brain signals from the different subjects will show a significant degradation in the accuracy as well as a higher complexity of the model generalisation. We assessed cross-subject speech detection according to the type of stimulus. EEG data were recorded in two sessions. In the first session the data were EEG data were acquired on the subjects who were tasked to pronounce the names of the colours displayed on the screen in front of them. In the second session, the subjects read the displayed text of the colours. Following the data acquisition, a comparison was carried out of the effect of these two types of visual stimuli on the speech detection. Nevertheless, the best result of speech detection was demonstrated on the data set from the reading session (F1 score of 83.69%, Table 3), the comparison of average results from all models showed that different stimuli did not have a significant effect on the detection results. The data from the colour naming session achieved a higher average F1 score just by 2.28% absolute, while the accuracy of both detections was at a very similar level. Therefore, we can assume that the speech detection system in our experiment did not have a high dependency on the nature/type of the stimulus, or that the EEG device used in the experiment did not have a sufficient resolution to distinguish between the type of the stimuli: image or text.

By combining the data sets from both sessions, a larger data set was obtained which was arranged in a number of different configurations which were then further analysed in the Neural Network. Table 4 shows that the best result was achieved for the configuration data set on subjects 1, 2 and, 3 in the training model and in the testing the model from the data on subject 4. The F1 score here reached 79.5%. Although a relatively small number of subjects was involved in the acquisition of the training data sets the high accuracy of the archived results provides an incentive for further research in this area.

Our results for the cross-subject speech detection achieved better results than initially anticipated. Based on our results we plan future work focused on the creation of an optimal model for the speech detection by involving a larger number of subjects and using deep machine learning algorithm. Such speech detection model could then be used to create a speech recognition application suitable for mobile EEG devices.

Author Contributions: Conceptualization, M.K.; methodology, M.K.; formal analysis, M.K., J.J.; investigation, M.K., J.J.; resources, M.K; writing—original draft preparation, M.K.; writing—review and editing, M.K. and J.J.; supervision, J.J.; project administration, J.J.; funding acquisition, J.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Slovak Research and Development Agency project number APVV-15-0517, the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic project number VEGA 1/0753/20 and the Cultural and Educational Grant Agency of the Slovak Republic project number KEGA 009TUKE-4-2019.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sharon, R.A.; Murthy, H.A. The "Sound of Silence" in EEG–Cognitive voice activity detection. *arXiv* **2020**, arXiv:2010.05497.
2. Dash, D.; Ferrari, P.; Dutta, S.; Wang, J. NeuroVAD: Real-Time Voice Activity Detection from Non-Invasive Neuromagnetic Signals. *Sensors* **2020**, *20*, 2248. [CrossRef] [PubMed]
3. Wang, J.; Kim, M.; Hernandez-Mulero, A.W.; Heitzman, D.; Ferrari, P. Towards decoding speech production from single-trial magnetoencephalography (MEG) signals. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 3036–3040.
4. Sharon, R.A.; Narayanan, S.S.; Sur, M.; Murthy, A.H. Neural Speech Decoding During Audition, Imagination and Production. *IEEE Access* **2020**, *8*, 149714–149729. [CrossRef]
5. Sereshkeh, A.R.; Trott, R.; Bricout, A.; Chau, T. Eeg classification of covert speech using regularized neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2292–2300. [CrossRef]
6. Krishna, G.; Tran, C.; Carnahan, M.; Han, Y.; Tewfik, A.H. Voice Activity Detection in presence of background noise using EEG. *arXiv* **2019**, arXiv:1911.04261.
7. Torres-Garcıa, A.A.; Moctezuma, L.A.; Molinas, M. Assessing the impact of idle state type on the identification of RGB color exposure for BCI. In Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (Biostec), Valletta, Malta, 24–26 February 2020.
8. Rosinová, M.; Lojka, M.; Staš, J.; Juhár, J. Voice command recognition using eeg signals. In Proceedings of the 2017 International Symposium ELMAR, Zadar, Croatia, 8 May 2017; pp. 153–156.
9. Koctúrová, M.; Juhár, J. Speech Activity Detection from EEG using a feed-forward neural network. In Proceedings of the 10th IEEE International Conference on Cognitive Infocommunications, Naples, Italy, 23–25 October 2019; p. 147.
10. Villena-González, M. The train of thought: How our brain responds to the environment whilst we are thinking in terms of mental images or an inner voice. *Cienc. Cogn.* **2016**, *10*, 23–26.
11. Breedlove, J.L.; St-Yves, G.; Olman, C.A.; Naselaris, T. Generative Feedback Explains Distinct Brain Activity Codes for Seen and Mental Images. *Curr. Biol.* **2020**, *30*, 2211–2224.e6. [CrossRef] [PubMed]
12. Winlove, C.I.; Milton, F.; Ranson, J.; Fulford, J.; MacKisack, M.; Macpherson, F.; Zeman, A. The neural correlates of visual imagery: A co-ordinate-based meta-analysis. *Cortex* **2018**, *105*, 4–25. [CrossRef] [PubMed]
13. Canini, M.; Della Rosa, P.A.; Catricalà, E.; Strijkers, K.; Branzi, F.M.; Costa, A.; Abutalebi, J. Semantic interference and its control: A functional neuroimaging and connectivity study. *Hum. Brain Mapp.* **2016**, *37*, 4179–4196. [CrossRef] [PubMed]
14. Biswas, S.; Sinha, R. Lateralization of Brain During EEG Based Covert Speech Classification. In Proceedings of the 2018 15th IEEE India Council International Conference (INDICON), Coimbatore, India, 16–18 December 2018; pp. 1–5.
15. Chakravarthy, V.S. A Gossamer of Words. In *Demystifying the Brain*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 321–347.
16. Koctúrová, M.; Juhár, J. Comparison of Dry Electrodes for Mobile EEG System. 2019. Available online: http://ceur-ws.org/Vol-2473/paper36.pdf (accessed on 30 November 2020).
17. Lamoureux, M.P.; Gibson, P.C.; Margrave, G.F. Minimum Phase and Attenuation Models in Continuous Time. 2011. Available online: https://www.crewes.org/ForOurSponsors/ResearchReports/2011/CRR201165.pdf (accessed on 30 November 2020).
18. Smith, A.D.; Ferguson, R.J. Minimum-phase signal calculation using the real cepstrum. *CREWES Res. Rep.* **2014**, *26*.
19. Bhakta, K.; Sikder, N.; Al Nahid, A.; Islam, M.M. Fault diagnosis of induction motor bearing using cepstrum-based preprocessing and ensemble learning algorithm. In Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, Bangladesh, 7–9 February 2019; pp. 1–6.
20. Agarwal, P.; Kale, R.K.; Kumar, M.; Kumar, S. Silent speech classification based upon various feature extraction methods. In Proceedings of the 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 27–28 February 2020; pp. 16–20.
21. Sanei, S.; Chambers, J.A. *EEG Signal Processing*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
22. Alías, F.; Socoró, J.C.; Sevillano, X. A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Appl. Sci.* **2016**, *6*, 143. [CrossRef]
23. Wolpaw, J.; Wolpaw, E.W. *Brain-Computer Interfaces: Principles and Practice*; OUP: New York, NY, USA, 2012.
24. Boubchir, L.; Daachi, B.; Pangracious, V. A review of feature extraction for EEG epileptic seizure detection and classification. In Proceedings of the 2017 40th International Conference on Telecommunications and Signal Processing (TSP), Barcelona, Spain, 5–7 July 2017; pp. 456–460.
25. Boashash, B.; Barki, H.; Ouelha, S. Performance evaluation of time-frequency image feature sets for improved classification and analysis of non-stationary signals: Application to newborn EEG seizure detection. *Knowl.-Based Syst.* **2017**, *132*, 188–203. [CrossRef]
26. Kiktova-Vozarikova, E.; Juhar, J.; Cizmar, A. Feature selection for acoustic events detection. *Multimed. Tools Appl.* **2015**, *74*, 4213–4233. [CrossRef]