Speech vs. Music Classification Based on EEG Spectral Features Using Artificial Neural Networks

Ivan Vajs, Predrag Jekić, Aleksandra Marjanović, Milica M. Janković

***Abstract*—The response mechanisms to different neural stimuli are a challenging task in neuroscience research. The auditory activity (response to music, speech, noise, etc.) can cause various emotional and cognitive responses. The neural responses to speech and music are of particular significance since they are almost constantly present in day-to-day life. We present the classification of the reactions to speech and music based on the spectral EEG features. The mean values of four frequency intervals (corresponding to the theta, alpha, beta, and gamma rhythms) were assessed for seven brain regions. These features were then used as the inputs to the classification based on logistic regression and artificial neural networks; both were used to analyze each subject individually and all available data. Feature selection was also performed, and the classification algorithms were trained using all, a half, and a quarter of the features for comparing their importance and variance for each individual and the entire dataset. The best classification accuracy for a single subject was 85.8%, and an accuracy of 67.1% was achieved for all subjects. This result is promising and calls for the analysis of a larger dataset.**

***Index Terms*—EEG; artificial neural networks; logistic regression; classification; feature selection.**

I. Introduction

The analysis of the neural responses to different stimuli is quite widespread in the neuroscience research [1]. One area of interest is the analysis of the relationship between the different types of auditory stimuli and brain activity. More specifically, speech and music are found to be of particular significance, as they are present in all cultures and play an important role in everyday life.

Different modalities can be used to track a person's neural response, with the three most commonly used being electroencephalography EEG [2], magnetoencephalography (MEG) [3] and functional magnetic resonance imaging fMRI [4]. These modalities measure the electrical neural activity, magnetic byproducts of neural activity, the changes related to blood flow, respectively, and are used in a wide range of studies regarding the functional analysis of the brain.

In [5], the authors explored the neural response of 15 subjects when exposed to, and when anticipating audio stimuli. The stimuli were selected to either be neutral, or to induce positive or negative emotions and the response was tracked using MEG. Each stimuli category was preceded with a cue tone of a different frequency so that the subject could know what emotion the following stimuli was meant to induce. It was shown that the brain response was different during the exposure to emotion inducing as opposed to neutral sounds, and that the response of a given stimuli was similar to the response elicited by its corresponding cue tone.

An investigation was carried out in [6] to determine whether neural separability between music and speech response could be detected. There were 47 participants that took part in the experiment and fMRI recordings were made during the exposure to short music excerpts and human vocalizations in a pseudo-random order. The results have shown that there is a specific brain region (a region within the anterior superior temporal gyrus) that responds more strongly to music than voice stimuli.

In [7], a study was conducted with the goal of classifying different musical notes based on the EEG response. Five participants took part in the experiment and the event-related spectral perturbation features were extracted and used as the input to the support vector machine classifier. The results of the study showed a 70% classification accuracy for 12 different classes (notes).

The classification of auditory stimuli (English vowels "a", "i" and "u") was conducted in [8]. Eight subjects took part in the experiment and a recurrent neural network combined with Ben's Spike Algorithm encoding was implemented to classify the EEG signals. The accuracy of 83.2% was obtained when using all 64 available electrodes, and an accuracy of 81.7% when using only 10 of the electrodes.

A classification of speech and music audio recordings was performed in [9]. Although not based on neural response, this paper is interesting because it implemented a novel Spectral Peak Tracking approach applied to the audio recording itself, to

[1]van Vajs is with the School of Electrical Engineering, University of Belgrade, and the Innovation Center, School of Electrical Engineering in Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: ivan.vajs@ic.etf.bg.ac.rs).

Predrag Jekić is with the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: jp205002p@student.etf.bg.ac.rs).

Aleksandra Marjanović is with the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: amarjanovic@etf.bg.ac.rs).

Milica M. Janković is with the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: piperski@etf.rs).

differentiate between music and speech. Very high classification accuracies (above 99%) achieved using deep learning techniques have shown that the complex structure of music and speech can be differentiated quite well.

Considering that there seems to be neural overlap between the brain response to music and speech [10], effectively distinguishing them using EEG could help separate these responses [11]. This can in turn, aid in the accuracy of the classification of the subjects focus point in the cases of exposure to multiple stimuli, which would be helpful in assistive therapies and the design of hearing loss devices [12].

Although the classification of audio stimuli has been attempted, no study was conducted to differentiate speech and music based on the spectral features of the subject's EEG, using artificial neural networks (ANN).

In this study, the classification of auditory stimuli into two categories (speech and music) was performed using a basic logistic regression model, as well as ANNs. The experiment setup and the EEG processing pipeline, along with the ANN architectures are described in Section II, the results are given in Section III, and the conclusion with directions for future work is given in Section IV.

II. Method

A. Experiment setup

Five healthy participants (Age: 31.4±8.8 years) took part in the experiment. All participants have signed the informed consent. A galvanic skin response (GSR) sensor (Mindfield Biosystems, Gronau, Germany) and an EEG cap (EASYCAP GmbH, Wörthsee, Germany) with 24 electrodes, placed in accordance with the 10–20 system connected to the 24-channel Smarting amplifier (mBrainTrain, Belgrade, Serbia) were used for the recording. Electrode M1 was used as an ECG channel and electrode M2 was excluded from the measurement in order to keep the symmetry of the EEG electrodes. In this study only the EEG signals were taken into consideration, with the sample rate of 256 Hz. The participants were asked to close their eyes and listen to the 30-minute-long audio file containing six sets of recordings trials. Each trial lasted four minutes with a one-minute-long interval of silence beforehand. Three types of audio recordings were played within a trial, each lasting for one minute, separated by 30-second silence intervals. One recording set consists of instrumental music, human speech and bird chirping. A single trial is a random permutation of the three mentioned recording categories. In this study, only the responses to the speech and instrumental music were analyzed. For two of the participants (ID2 and ID3) the measurements from the final third of the experiment were excluded due to the reported discomfort of the participants.

B. EEG processing

Firstly, the recorded EEG signals were filtered using a notch filter to remove the power supply noise at 50 Hz. The EEG corresponding to the music and speech stimuli was cut into data snippets using a time window of two seconds and the time stride of two seconds (i.e., non-overlapping time windows). The data snippet was labeled according to its corresponding stimuli. The feature extraction process was performed on each snippet and consists of the following steps. An estimation of the power spectral density (PSD) was performed electrode-wise, denoted as PSD_e (PSD for electrode e). For every PSD_e, a reference $PSD_{e,ref}$ was extracted from the 10-second interval of silence which precedes the particular stimulus recording. The difference between PSD_e and its respective $PSD_{e,ref}$ (denoted as $PSD_{e,diff}$) was then calculated. At this point, the 22 observed EEG channels (their corresponding $PSD_{e,diff}$) were grouped into 7 categories as follows:

- Frontal left: Fp1, F3, F7, Fz, AFz;
- Frontal right: Fp2, F4, F8, Fz, AFz;
- Central: C3, C4, Cz, CPz;
- Parietal left: P3, P7, Pz, POz;
- Parietal right: P4, P8, Pz, POz;
- Occipital: O1, O2;
- Temporal: T7, T8.

The electrodes were grouped according to their position (frontal, central, parietal, occipital, and temporal), with the frontal and parietal regions being split into two hemispheres, since the number of electrodes in each of the hemispheres was sufficient for them to be observed independently. The $PSD_{e,diff}$ of the electrodes in a single category were averaged, thus creating seven $PSD_{g,diff}$ (PSD for group g, $g \in 1 \div 7$). Finally, the mean spectral power of the following frequency bands (i.e., brainwave activity [13]) was estimated for each $PSD_{g,diff}$:

- [4 Hz, 8 Hz] – theta;
- [8 Hz, 12 Hz] – alpha;
- [12 Hz, 30 Hz] – beta;
- [30 Hz, 80 Hz] – gamma.

This resulted in 4 frequency bands × 7 groups = 28 features for the classification algorithms.

C. Classification algorithms

Multiple models were engineered for the purposes of this study and evaluated using 20-fold cross-validation [14]. In the first part of the study, both a logistic regression (used as a baseline algorithm) and an ANN architecture were designed and trained per participant. The same architectures were used with either 7, 14 or all 28 features as inputs. In the cases of 7 and 14 chosen features, the selection was based on the ANOVA F-value estimated on the training set (Fig. 1.) [15].

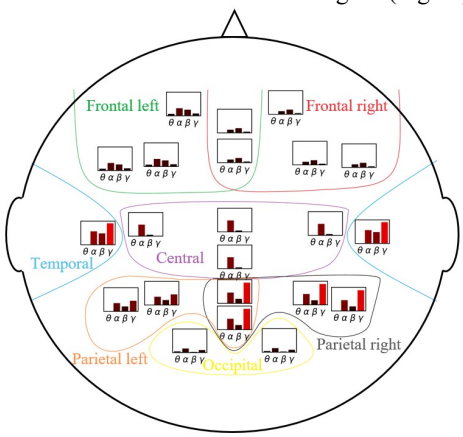

Fig. 1. F-value visualized for each electrode group for a single participant, plotted on the locations of all observed EEG electrodes.

To summarize, there were two different model architectures, three different numbers of input features and five different participants, adding up to 30 different models in total, that were evaluated in this part of the study. In the second part of the study, one logistic regression and three ANN architectures were designed and trained on all collected data, i.e., the data from all participants was placed into a single dataset. These architectures were used for building up different models with either 7, 14 or all 28 features as inputs. Having had four different model architectures and three different numbers of input features, this meant building up 12 different models in total, in the second part of the study.

The shallow ANN used in both the first and second part of the study (ANN1) contained three fully connected (FC) layers (with 5, 4, 2 neurons, respectively). The first hidden layer of ANN1 was a random projection layer, with the purpose of adaptive dimensionality reduction [16]. The other two ANN architectures (ANN2 and ANN3) were evaluated only in the second part of the study. ANN2 contained three FC layers (with 5,10, 2 neurons, respectively). ANN3 consisted of four FC layers (with 5, 15, 10, 2 neurons, respectively). Both ANN2 and ANN3 had a random projection layer as their first hidden layer for the same reason as ANN1. The ANN architectures used for all 28 input features are shown in Fig. 2 (the varying number of input features changes the size of the input layer).

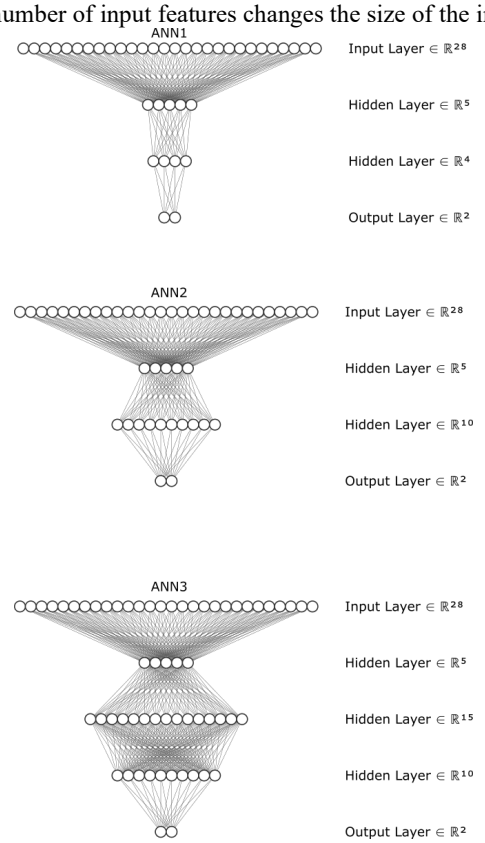

Fig. 2. ANN architectures for all 28 input features.

The relatively low number of neurons and layers was chosen to avoid overfitting considering the dataset size. Adam optimizer was used for the training of the networks with the initial learning rate of 0.00005, batch size of 8, 350 epochs and leaky ReLU activation functions for all hidden layers [17], [18].

III. Results

In Table I, the test classification accuracies are listed per architecture for each subject.

TABLE I
CLASSIFICATION ACCURACY [%] FOR EACH SUBJECT.

	Logistic regression			**ANN1**		
	Number of features			**Number of features**		
ID	28	14	7	28	14	7
1	82.9	78.7	78.4	85.8	81.0	79.7
2	74.5	73.3	65.1	75.2	74.2	67.4
3	81.1	75.1	68.4	84.7	76.2	69.5
4	75.6	72.5	69.5	77.6	73.9	70.1
5	74.3	73.4	71.7	74.9	74.3	72.3

The obtained results show that a higher number of input features corresponds to the higher accuracy regardless of the subject and algorithm (which stands in line with the results from [8]). Furthermore, for a given subject and number of input features, ANN1 consistently achieves a higher accuracy compared to the logistic regression. It is important to note that each subject has their specific features that are consistently selected throughout the cross-validation folds and that these features vary between the subjects (Table II). This is expected due to the natural variation of EEG responses between individuals [19].

TABLE II
THE SELECTED FEATURES FOR EACH PARTICIPANT SHOWN FOR THE TWO FOLDS THAT EXHIBIT THE BIGGEST DIFFERENCE IN FEATURE SELECTION [GROUP NUMBER FOLLOWED BY BRAINWAVE SYMBOL].

ID	**fold**	**14 selected features**	**7 selected features**
1	1	1α, 1β, 2α, 2β, 3α, 4α, 4β, 4γ, 5α, 5β, 5γ, 7α, 7β, 7γ	1α, 3α, 5α, 5γ, 7α, 7β, 7γ
	2	1α, 1β, 2β, 3α, 4α, 4β, 4γ, 5α, 5β, 5γ, 6α, 7α, 7β, 7γ	3α, 4γ, 5α, 5γ, 7α, 7β, 7γ
2	1	2α, 2γ, 3θ, 3α, 3β, 4θ, 5β, 5γ, 6θ, 6α, 6β, 6γ, 7θ, 7γ	2γ, 3θ, 3α, 3β, 5β, 6α, 7γ
	2	1β, 2γ, 3θ, 3α, 3β, 3γ, 5β, 5γ, 6α, 6β, 6γ, 7θ, 7α, 7γ	3θ, 3α, 3β, 4θ, 5β, 6α, 7γ
3	1	1θ, 1β, 2θ, 2α, 2β, 3θ, 3α, 3β, 4β, 4γ, 5α, 5β, 5γ, 6β	3α, 3β, 4β, 4γ, 5β, 5γ, 6β
	2	2θ, 2α, 2β, 3θ, 3α, 3β, 4β, 4γ, 5α, 5β, 5γ, 6θ, 6β, 7α	2θ, 3θ, 3α, 3β, 4γ, 5β, 5γ
4	1	1β, 1γ, 2θ, 2β, 2γ, 3γ, 4α, 4γ, 5α, 5β, 5γ, 6α, 6γ, 7α	1γ, 2γ, 3γ, 4α, 4γ, 5γ, 6α
	2	1β, 1γ, 2θ, 2β, 2γ, 3α, 3γ, 4α, 4γ, 5α, 5γ, 6α, 6γ, 7α	1γ, 2γ, 3γ, 4γ, 5γ, 6α, 6γ
5	1	1θ, 1α, 1γ, 2θ, 2α, 2γ, 3β, 4θ, 4β, 5β, 5γ, 6β, 6γ, 7γ	1θ, 1γ, 2θ, 2γ, 4β, 5γ, 6β
	2	1θ, 1α, 1γ, 2θ, 2α, 2γ, 4θ, 4β, 4γ, 5β, 5γ, 6β, 6γ, 7γ	1θ, 1γ, 4β, 5β, 5γ, 6β, 6γ

In Table III, the test classification accuracies are listed for each model deployed on the set containing the data from all the subjects.

TABLE III
CLASSIFICATION ACCURACY [%] FOR ALL SUBJECTS.

	Number of features		
Architecture	28	14	7
Logistic regression	61.4	59.1	58.3
ANN1	64.8	63.3	62.1
ANN2	65.6	64.2	63.0
ANN3	67.1	64.8	63.9

The overall accuracies shown in Table III are lower than the accuracies obtained when the individual subjects were considered. With respect to the diversity of individual EEG responses and the number of participants it was more difficult for the algorithms to pick up on the complex input-output dependencies.

IV. CONCLUSION

In this paper, the classification of audio stimuli (speech and music) based on spectral EEG features was performed. Firstly, the classification was performed per subject, using the logistic regression and ANN. ANN has shown a slight but consistent improvement (ANN accuracy ranging from 67.4% to 85.8%) over the baseline logistic regression which is to be expected considering the dataset size. Furthermore, a larger number of input features implies a small but consistent increase in accuracy. The deployed models achieved an accuracy above 65% on the test set even when 7 features were selected from the observed dataset. This implies that although a higher number of input features does improve the overall accuracy, certain features do carry more useful information than others. On the other hand, having all collected data in one dataset, resulted in having the maximum accuracy of 67.1%. This is due to the difficulty of achieving higher accuracies when there is an undeniable diversity in the dataset compared to the number of instances and a varying importance of a single feature between subjects.

The directions for the future work include expanding the dataset with significantly more subjects, thus enabling the development of more complex algorithms, alongside the implementation of other EEG processing and feature selection methods. By expanding the database and expanding the EEG feature set, a higher distinction accuracy between speech and music response could be expected. This would open up a possibility to estimate the focus of a given subject when exposed to these stimuli simultaneously, which is often the case in day-to-day life. Further research will also include emotional aspects based on the consideration of heart rate variability (HRV) parameters and the GSR.

ACKNOWLEDGMENT

This research was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia. The authors would also like to thank all the participants that took part in the experiment and dr Andrej Savić for his useful advice regarding the processing of the EEG signals.

REFERENCES

[1] A. F. Meyer, R. S. Williamson, J. F. Linden, and M. Sahani, "Models of neuronal stimulus-response functions: Elaboration, estimation, and evaluation," *Frontiers in Systems Neuroscience*, vol. 10. Frontiers Media S.A., p. 109, 12-Jan-2017.

[2] C. D. Binnie and P. F. Prior, "Electroencephalography," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 57, no. 11. BMJ Publishing Group, pp. 1308–1319, 1994.

[3] S. P. Singh, "Magnetoencephalography: Basic principles," *Ann. Indian Acad. Neurol.*, vol. 17, no. SUPPL. 1, p. S107, 2014.

[4] G. H. Glover, "Overview of functional magnetic resonance imaging," *Neurosurgery Clinics of North America*, vol. 22, no. 2. NIH Public Access, pp. 133–139, Apr-2011.

[5] K. Yokosawa, S. Pamilo, L. Hirvenkari, R. Hari, and E. Pihko, "Activation of auditory cortex by anticipating and hearing emotional sounds: An MEG study," *PLoS One*, vol. 8, no. 11, p. 80284, Nov. 2013.

[6] J. L. Armony, W. Aubé, A. Angulo-Perkins, I. Peretz, and L. Concha, "The specificity of neural responses to music and their relation to voice processing: An fMRI-adaptation study," *Neurosci. Lett.*, vol. 593, pp. 35–39, Apr. 2015.

[7] K. Tsekoura and A. Foka, "Classification of EEG signals produced by musical notes as stimuli," *Expert Syst. Appl.*, vol. 159, p. 113507, Nov. 2020.

[8] M.-A. Moinnereau, T. Brienne, S. Brodeur, J. Rouat, K. Whittingstall, and E. Plourde, "Classification of auditory stimuli from EEG signals with a regulated recurrent neural network reservoir," Apr. 2018.

[9] M. Bhattacharjee, S. R. M. Prasanna, and P. Guha, "Speech/Music Classification Using Features From Spectral Peaks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1549–1559, 2020.

[10] I. Peretz, D. Vuvan, M.-É. Lagrois, and J. L. Armony, "Neural overlap in processing music and speech," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 370, no. 1664, p. 20140090, 2015.

[11] N. J. Zuk, E. S. Teoh, and E. C. Lalor, "EEG-based classification of natural sounds reveals specialized responses to speech and music," *Neuroimage*, vol. 210, p. 116558, Apr. 2020.

[12] M. Ogg, D. Moraczewski, S. E. Kuchinsky, and L. R. Slevc, "Separable neural representations of sound sources: Speaker identity and musical timbre," *Neuroimage*, vol. 191, pp. 116–126, May 2019.

[13] C. S. Nayak and A. C. Anilkumar, *Eeg normal waveforms*. StatPearls Publishing, 2020.

[14] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *J. R. Stat. Soc. Ser. B*, vol. 36, no. 2, pp. 111–147, Jun. 1974.

[15] L. Sthle and S. Wold, "Analysis of variance (ANOVA)," *Chemometrics and Intelligent Laboratory Systems*, vol. 6, no. 4. Elsevier, pp. 259–272, 01-Nov-1989.

[16] P. Iwo, Wojcik, "Random Projectionin Deep Neural Networks," University of Science and Technology in Kraków, 2018.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv Prepr. arXiv1412.6980*, 2014.

[18] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv Prepr. arXiv1505.00853*, 2015.

[19] J. Intriligator and J. Polich, "On the relationship between EEG and ERP variability," *Int. J. Psychophysiol.*, vol. 20, no. 1, pp. 59–74, Jun. 1995.