Fatigue Analysis Based On Voice Spectrogram Images Using Deep Learning

Uğur Keskin, Hamza Osman Ilhan
Department of Computer Engineering
Yildiz Technical University, 34220 Istanbul, Türkiye
l1117044@std.yildiz.edu.tr

***Özetçe* —Yorgunluk, gündelik hayatta çeşitli kazalara ve hatalara yol açabilen yaygın bir durumdur. Bu kazalara özellikle trafik kazaları, iş kazaları örnek verilebilir. Bu kazaları ve hataları önlemek için yorgunluk durumunun oluşması durumunda kaza yapma potansiyeline sahip olan yorgun bireyi uyarmak ya da genel anlamda bu konuda veri toplayıp bunları iş saatlerini düzenlemekte kullanmak gibi sistemler mevcuttur.**

Yorgunluk, insanın anlık yüz özelliklerine ya da ani olaylara verdiği tepki süresinden anlaşılabildiği gibi, insanın sesinden de anlaşılabilir. İnsan sesinin nicel olarak nasıl ifade edilebileceği konusunda da farklı fikirler mevcuttur.

Bu projede, bir derin öğrenme metodu olan CNN modeli kullanılarak insan seslerinin ne kadar yorgunluk belirttiğine dair tahminler yapılmıştır. Bu tahminlerin yapılırken, insan sesinin nicel verilerle ifade edilebilmesi için, seslerin spektrogram görüntülerinin renk değerleri kullanılmıştır. Proje sonucunda ses spektrogramlarının dinç ya da yorgun olarak sınıflandırılmasında yaklaşık %94'lük bir doğruluk oranına ulaşılmıştır.

Anahtar Kelimeler—Yorgunluk ses analizi, ses spektrogramları, transfer öğrenme, derin öğrenme, yapay sinir ağları, ses sınıflandırma

***Abstract*—Fatigue is a common condition that can lead to various accidents and mistakes in everyday life. Traffic accidents, occupational accidents can be given as examples. In order to prevent these accidents and errors, there are systems such as to alert the tired individual who has the potential to have an accident in case of fatigue or to collect data in this regard and to use them to organize the working hours.**

As well as fatigue can be understood from a person's reaction time to sudden events or from her/his current facial features, it can be understood from a person's voice. Also there are plenty options for expressing human voice as quantitative data.

In this project, by using the CNN model, a deep learning method, estimates of how tired the human voices indicate are made. In making these predictions, the color values of the spectrogram images of the voices were used to express the human voice with quantitative data. As a result of the project, an accuracy rate of approximately %94 has been reached in the classification of sound spectrograms as vigorous or tired.

Keywords—Fatigue voice analysis, voice spectrograms, transfer learning, deep learning, neural networks, audio classification

I. Introduction

Fatigueness which often causes inattentiveness, inefficiency and errors on work is an important factor in life. Fatigue can be achieved through certain movements and expressions of people. However, long driving time is an important cause of fatigue accidents on highways or major highways.[1], For this reason, fatigue that emerges during the day is an important factor causing traffic accidents and fatal work accidents. One of the factors that indicates this is the voice of the human. When people are tired, this fatigue may be reflected, so it is possible to understand if he/she is tired from the human voice. The application has developed for those who need fatigue analysis over voice in their own research.

In this project, deep learning method has used and the most successful deep learning model was tried to be developed by optimizing the selected CNN structures in order to analyze whether the person owning the sound is tired by making use of the sound spectrogram images obtained with different techniques. Also, an application with an interface has been developed that will allow the user to enter the voice spectrogram image that she/he wants to classify and inform the user which class it belongs to with the help of the developed deep learning model.

II. Previous Works

In study at [2], estimation of human fatigue is targeted using sound data. In this study, which was carried out using sound analysis methods, 296 speeches from 31 normal people divided into 3 different groups were recorded and used as data. After these sound recordings were passed through the Fourier Transform, they were scaled according to the response characteristics of the human ear with the Mel-scale Filterbank Analysis method and 36 different variables were obtained for each frame of the voice data by passing through 2 derivative processes together with Cepstral analysis. VCM (Sound Correlation Metric) metrics in which the words SOL (Sleep Delay Test) and p (in the word "pea") and t (in the word "tea") are tried to be labeled by scoring sound data as sleepy-sleepless. Error between predicted value and real value has calculated by using R^2 and as a result, predicted values reached %68 accuracy for SOL metric, %79 accuracy for VCM("p") and %45 accuracy for VCM("t").

In the study for drivers at [3], An application that detects the fatigue of the driver has been developed by monitoring the facial expressions of the driver and estimating the images obtained. In the study, which was analyzed separately for different parts of the face, eye movement, pupil movement, head direction prediction, eye gaze direction detection, facial expression analysis were used. The data obtained from the facial movements benefited are classified by the BN method and it is explained how fatigue appears in each state of each data and how likely. With classification by looking the states when average winking speed is high and flexing frequency is low , model got %95.52 accuracy score.

In study at [4] , sound recording of 12 participants whose sound recordings were taken was examined in detail, The voice recordings of 2 participants were analyzed in pieces due to sound problems. 1 hour just before each voice recording of the participants, who were kept from sleep, was scored between 1 and 10 according to fatigue using the CSR method. The sound recordings obtained were separated into acoustic properties with Praat Speech Analysis software, thus, a total of 45,216 variables were obtained for each speech sample. The data obtained was tried to be classified by giving separate information to 8 machine learning methods. That methods are MLP, SVM_l, SVM_r, 5-NN, Decision Tree, PC, Logistic Regression ve LDA methods. To achieve a better result, an ensemble classified model was created by averaging the classification outputs of Machine Learning methods. The EC model formed was tested with sound data not previously used in the training of the model, the success rate obtained was recorded as %83.

As it is understood from previous studies, in this project, unlike previous studies, I aimed to provide high performance by using different data type and different solution method. The data I used are sound spectrogram images that are formed by converting the sound with various methods, the solution of fatigue detection is tried to be found by using the deep learning method trained using tagged data.

III. MATERIALS AND METHODS

A. Method

A deep learning structure known as Convolutional Neural Networks were used to classify voice spectrogram images of voice which belongs to the image is a fatiguous voice or not. I used transfer learning to get pre-trained convolutional neural network models which trained with "imagenet" dataset for getting more successfull results.

The data required for training the established deep learning model are voice spectrogram images. Images were created separately for each of the 4 spectrographic types, BARK, ERB, LOG, MEL in MATLAB environment. These images were labeled as tired or vigorous, with more than one person scoring between 1-10, and different folders were created according to their scores.

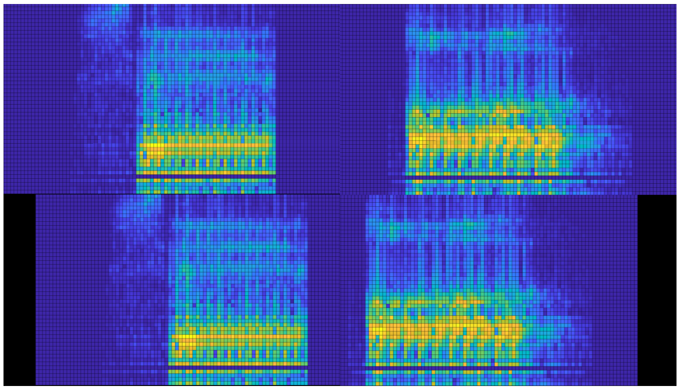
Each of the different folders created for score ranges with each spectrogram creation method used contains 5 folders in order to comply with the 5-fold validation architecture to make the training more consistent and accurate.

Pre-training a CNN model using ImageNet first, and then retraining it to adapt to a new goal has become a standard for solving computer vision problems. In order to find out which CNN Model will reach the solution better, the scores of 7 artificial neural network models determined on this project have been examined. These are Resnet50, VGG16, VGG19, MobileNet, MobileNetV2, DenseNet201 and NasNetMobile models.

Many optimization functions have been tried to give Deep Learning training more accurate and consistent results, these are Adam, SGD, RmsProp, Nadam, Adamax and Adadelta functions. In addition to these functions, it is aimed to determine the variables such as learning rate coefficient, number of trainable hidden layers in convolutional neural networks, activation functions to be determined for each extra layer, epoch number of training, batch size and loss function.

In the developed CNN model, the activation function of the final layer was determined as softmax, with the formula $S(y_i) = \frac{exp(y_i)}{\sum_j exp(y_j)}$, and the error function was determined as Categorical Cross Entropy with the formula $CCE(p,t) = -\sum t_{o,c} \log p_{o,c}$.

In order to ensure that the data can be used for training and testing, the dimensions of the images to be used for training and testing are RGB type and RGB values normalized between 0 and 1, converted from 224x224x3 pixel sizes from 1800x1013x3 pixel sizes to match the input layer of the neural networks used. normalized. The data are tagged using the "one hot encoding" method.



(a) **(b)**

Figure 1 Horizontally Scrolled Sound Spectrogram at Certain Proportions

One of the methods used to prevent the data from trying to estimate the data by fitting a wrong pattern and to reduce the problem of overfitting is to augment or diversify the image by performing operations such as whitening, mirroring, scrolling, and rotating on the images. However, it is necessary to take into account the possibility that this formatting process may lose the meaning of the images. With this reason, it has been tried to format the images with only horizontal scrolling in the voice spectrogram images. Parameters that determine how much the image will be shifted and how to fill the gap after shifting are also variables that affect the optimization of the training of the network. Visual 1 shows examples of an voice spectrogram image formatted by horizontal shifting, "a" tag specifies fatiguous human voice spectrogram, "b" tag specifies vigorous human voice.

Number Of Sample		Fatigueness Level								
		1-10			1,2-9,10			1,2,3-8,9,10		
		Vigorous	Fatiguous	Total	Vigorous	Fatiguous	Total	Vigorous	Fatiguous	Total
Letter	BARK	32	87	119	131	197	328	438	396	834
	ERB	32	87	119	131	197	328	438	396	834
	LOG	32	87	119	131	197	328	438	396	834
	MEL	32	87	119	131	197	328	438	396	834
Word	BARK	40	112	152	176	254	430	497	428	925
	ERB	23	60	83	175	251	426	596	519	1115
	LOG	41	112	153	171	253	424	597	518	1115
	MEL	43	105	148	173	261	424	584	525	1109
Combine	BARK	72	199	271	307	451	758	935	824	1759
	ERB	55	147	202	306	448	754	1034	915	1949
	LOG	73	199	272	302	450	752	1035	914	1949
	MEL	75	192	267	304	458	752	1022	921	1943

Table 1 Number of samples grouped by spectrogram type, voice type, fatigueness level and class

To determine which of the Resnet50, VGG16, VGG19, MobileNet, MobileNetV2, DenseNet201 and NasNetMobile convolutional neural networks will be focused on, each network's error function is selected as Categorical Cross Entropy, accuracy metric selected as test metric for measuring the success of the network and softmax function selected as networks' function at their last layer. accuracy was taken into account, 6 different optimization functions were tried for each network and test success rates were recorded throughout the training.

The optimization functions used are Nadam, Adam, Adamax, Adadelta, RMSProp and SGD. Training and test The images in the 5 folds, which has 265 voice spectrogram images for training and 65 voice spectrogram images for each test on the MEL typed and 1,2,9 or 10 scored dataset Trained with and tested with images in its own test data, success rates were determined as the average of the highest test success achieved by these 5 folds. The neural network-optimization function pair, which reached the highest accuracy rate, was selected to be used in subsequent optimization processes.

To measure the success of the model, for each folder prepared according to the 5 fold architecture, the model was trained with the training folder of each fold of that folder and tested with the test folder of the same fold. To record the success of the folder, the average of the test values of all the folds of the folder and the highest value reached were taken into account. The error function used for the test is the CCE (Categorical Cross-Entropy) function and the accuracy metric is used as the metric.

B. *Dataset Information*

The latest model formed is 1 or 10; 1,2,9 or 10; It was individually trained and tested on databases containing spectrogram images of sounds rated as 1,2,3,8,9 or 10, and these images were emphasized while performing success assessments. Success and Error percentages are expressed by showing the average of 5 folds of folders and the value of the highest fold reached.

As can be seen in Table 1, in the dataset consisting of spectrogram images of letter sounds scored as 1 or 10, 96 images for training and 23 images for test; The data set consisting of spectrogram images of word sounds scored as 1 or 10 contains 67 images for training, 16 images for testing and the combination of these datasets, 163 images for training and 39 images for testing.

In another dataset consisting of spectrogram images of letter sounds scored as 1,2,9 or 10, 265 images for training and 65 images for testing; The data set consisting

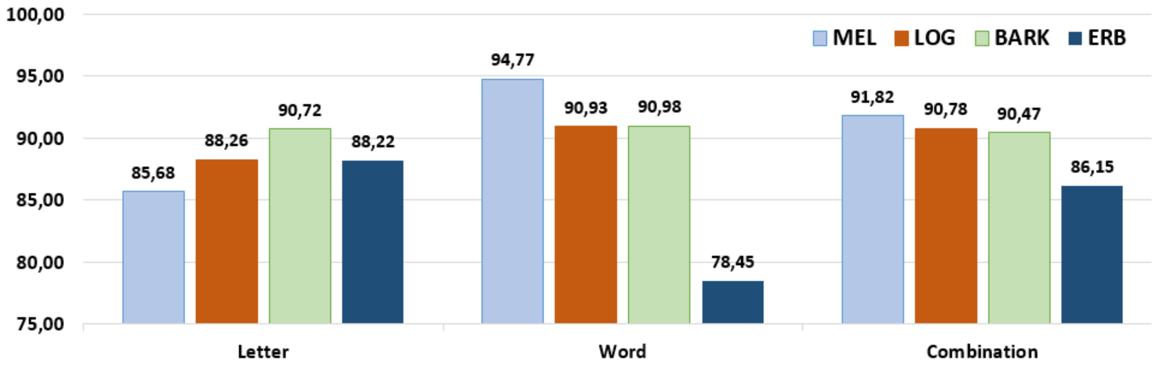


Figure 2 Test results on voice datasets scored as 1 or 10

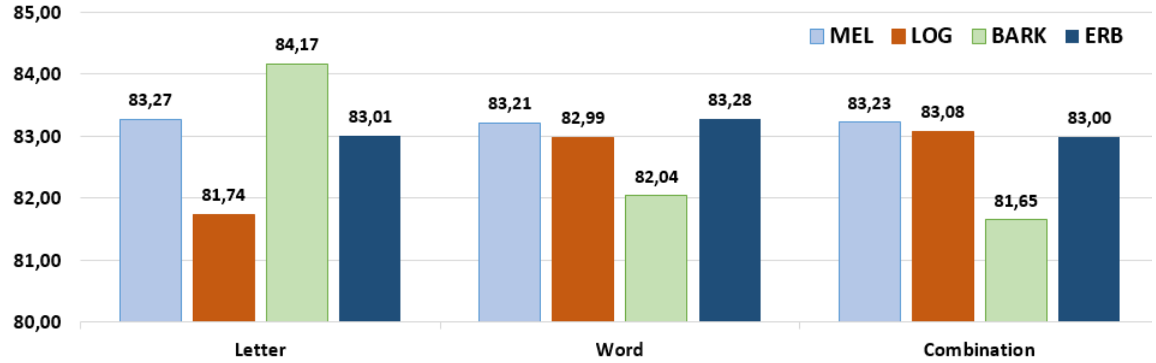


Figure 3 Test results on datasets scored as 1, 2, 9 or 10

of spectrogram images of word sounds scored as 1,2,9 or 10 contains 594 images for training, 124 images for testing and 757 for training and 189 images for testing in the combination of these datasets.

In another dataset consisting of spectrogram images of letter sounds rated as 1,2,3,8,9 or 10, 667 for training and 167 for test; The data set consisting of spectrogram images of word sounds scored as 1,2,3,8,9 or 10 contains 494 images for training, 227 images for testing and 1161 for training and 394 images for testing in the combination of these datasets.

Accuracy for each spectrogram type consists "MEL", "BARK", "ERB" and "LOG" are shown in following figures. Figures with "a" tag specifies letter voice datasets, "b" tag specifies word voice datasets and "c" tag specifies combined voice datasets.

IV. Experimental Results

In the measurements made, VGG16 has shown the highest performance among the tested models to find the best transfer learning model that can provide the solution of the problem being handled, and after the optimization attempts on this model, the latest model was trained and tested on different images produced with different spectrogram types.

As shown in figures at Figure 2, MEL type spectrogram images were the highest success among all folders, with %94.77 in test results on word folders. In addition, the highest fold score achieved was obtained with MEL-type spectrogram images with %100 .

As shown in figures at Figure 3 figures, BARK type spectrogram images with %84.17 in the test results on word folders achieved the highest success among all folders, as shown in the figures. In addition, the highest fold score achieved was obtained with BARK type spectrogram images with %93.84 .

As shown in figures at Figure 4, MEL type spectrogram images were the highest success among all folders with %80.63 in test results on word folders. In addition, the highest fold score achieved was obtained using LOG type spectrogram images with %84.37 .

According to the results obtained, assuming 1, 2, 9 and 10 scored voices' spectrograms are most accurate data, BARK spectrogram type and letter sounds have the highest accuracy rate with 84.17 fold average and % 93.84 highest one fold success. When the accuracy matrices of the models are examined, it is concluded that the models can generally distinguish between tired human voices more easily than vigorous human voices.

Figure 5 contains 3 confusion matrix of best test results on letter, word and combined voice types respectively. The most successful spectrogram type being classified when converted from letter voice types is BARK whose confusion

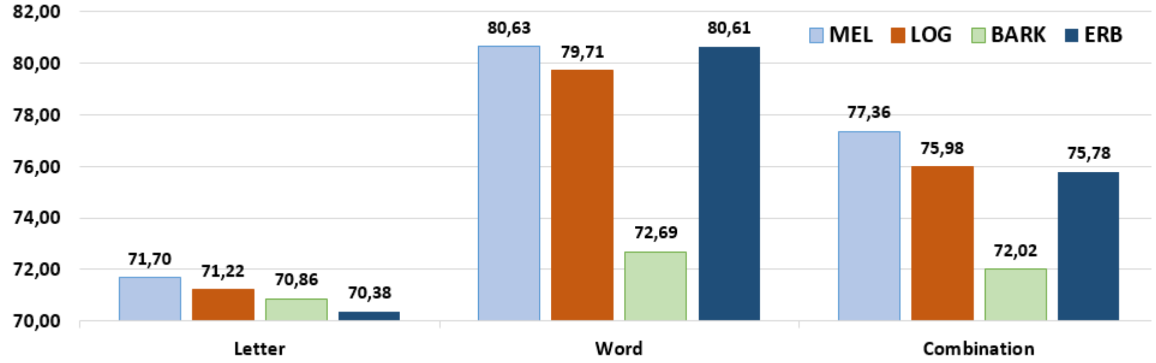
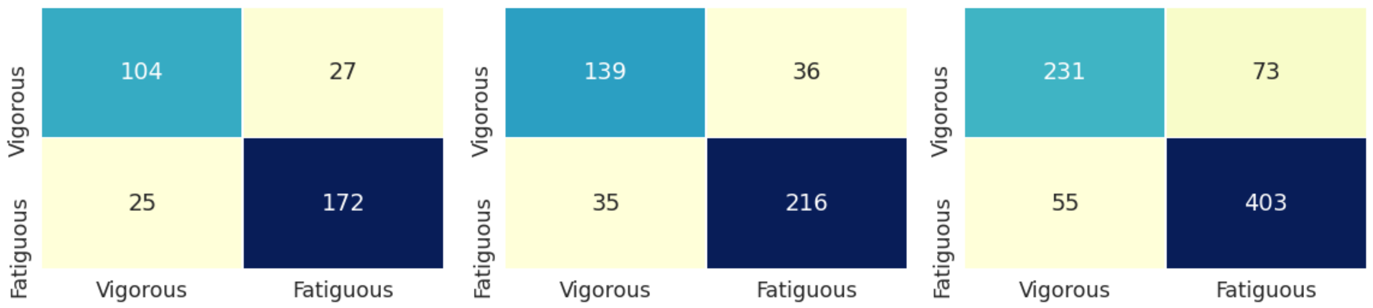


Figure 4 Test results on datasets scored as 1, 2, 3, 8, 9 or 10



(a) (b) (c)

Figure 5 Confusion Matrices Of Best Test Results on Letter, Word and Combined Voice Type Respectively

matrix is shown in (a), The most successful spectrogram type being classified when converted from word voice types is ERB whose confusion matrix is shown in (b) and The most successful spectrogram type being classified when converted from combination of both letter and word voice types is MEL whose confusion matrix is shown in (c).

As can be understood from the confusion matrix based performance metrics on Table 2, deep learning model trained by letter voice spectrogram datasets are getting slightly more accurate results at classifying fatigueness. Test results on all datasets are getting higher precision, recall and f-score on classifying fatiguous voices than vigorous voices.

V. Conclusion

In this project, the classification of the sound based on its fatigueness level was made by classifying the sound spectrogram images obtained by converting the sound into 4 different spectrogram types MEL, ERB, BARK and LOG unlike previous works and researchs on fatigue analysis. In this classification problem, CNN models were used as a solution . In order to make the model more successful, 7 CNN structures which previously trained using ImageNet datasets, were adapted to the problem by using transfer learning.

In order to optimize on these networks, a pre-test was performed between them and the model to be fine-tuned was determined. Among these networks, the VGG16 was the network that best adapted to the problem. Various optimizations on dataset and network were made to improve accuracy of classification. These optimizations include image augmentation, activation functions, additional layers, hyperparameter optimizations, metric types and optimizer functions.

Results of classification on BARK spectrogram images which converted from letter voices have the highest accuracy among other spectrogram types converted from letter voices, furthermore results of ERB spectrogram images converted from word voices have the highest accuracy among other spectrogram types converted from word voices and results of MEL spectrogram images converted from a combined dataset which contains spectrogram images which converted from both word and letter voices have the highest accuracy converted from combined dataset. In addition, a software that classifies the spectrogram image using saved convolutional neural network model entered according to the spectrogram type and voice type that spectrogram converted from selected by the user has been developed for making predictions on new images collectively or singularly.

References

[1] P.-H. Ting, J.-R. Hwang, J.-L. Doong, and M.-C. Jeng, “Driver fatigue and highway driving: A simulator study,” *Physiology & behavior*, vol. 94, no. 3, pp. 448–453, 2008.

[2] H. P. Greeley, J. Berg, E. Friets, J. Wilson, G. Greenough, J. Picone, J. rey Whitmore, and T. Nesthus, “Fatigue estimation using voice analysis,” *Behavior research methods*, vol. 39, no. 3, pp. 610–619, 2007.

[3] Z. Zhu and Q. Ji, “Real time and non-intrusive driver fatigue monitoring,” in *Proceedings. The 7th International IEEE Conference on Intelligent Transportation Systems (IEEE Cat. No. 04TH8749)*. IEEE, 2004, pp. 657–662.

[4] J. Krajewski, U. Trutschel, M. Golz, D. Sommer, and D. Edwards, “Estimating fatigue from predetermined speech samples transmitted by operator communication systems,” 2009.

	Letter		Word		Combination	
	Vigorous	Fatiguous	Vigorous	Fatiguous	Vigorous	Fatiguous
Precision	0.79	0.87	0.79	0.86	0.76	0.88
Recall	0.81	0.86	0.80	0.86	0.81	0.85
F-Score	0.80	0.87	0.80	0.86	0.78	0.86
Accuracy	0.841		0.833		0.832	
Kappa	0.669		0.655		0.646	

Table 2 The evaluation of best classifiers in terms of confusion matrix based performance metrics derived from Figure 5