DOI https://doi.org/10.15407/csc.2021.04.028
UDC 004.032.26

A.V. MANOKHIN, student, Computer Systems Software Department of the Applied Mathematics Faculty National Technical University of Ukraine "Igor Sikorsky Kyiv Politechnic Institute",
Peremohy Ave., 37, Kyiv, Ukraine 03056,
manokhin.kpi@gmail.com

N.A. RYBACHOK, PhD (Eng.), Senior Lecturer, National Technical University of Ukraine "Igor Sikorsky Kyiv Politechnic Institute",
Peremohy Ave., 37, Kyiv, Ukraine, 03056,
rybachok@pzks.fpm.kpi.ua

ENGLISH ACCENT RECOGNITION USING DEEP MACHINE LEARNING

The article highlights aspects of the use of deep machine learning to recognize the accents of the English language. Software has been developed to determine the percentage of how close audio recordings are to each of 8 most common English accents. A convolutional neural network consisting of 2 convolutional layers, 1 max pooling layer, and 2 dense layers was trained across 2 epochs on a set of 5,516 audio recordings taken from the English Multi-speaker Corpus for Voice Cloning resource. The forecasting accuracy of 89,07% was achieved on the test data presented by 11 thousand MFCC matrices with a dimension of 50x87.

Keywords*: neural network, deep machine learning, accent recognition, MFCC, CNN.*

Introduction and Problem Statement

Considering drastic surge of ways to communicate nowadays, many people are in need to use popular international languages. The most popular in fact is the English language which is used by 1,348 bil. people [1].

In modern digitalized world ASR (Automatic Speech Recognition) systems propose new interfaces for human-machine communication. Not only it facilitates convenience for this means of communication, but also widens potential user base of this software, for example, people with disabilities.

However, another crucial problem of English accents recognition emerges [2]. For regular people it is described by their demand in learning the language or its particular accent. ASR systems developers strive to improving their systems thus acknowledging users' accent. It is possible to use accent recognition for immigrational screening [3], in video games [4] and recommendation systems. The aim of this work is to develop a neural network and according software for English accent recognition.

The following tasks need to be done:

- determine the requirements of the neural network and developing software;
- develop a neural network and appropriate software;
- analyze the work of the neural network and software.

Literature Overview

Neural networks and deep machine learning are often used to recognize and classify audio and video objects [2–5]. The best results of predictions in the analysis of human speech are demonstrated

by neural networks using MFCC (Mel-Frequency Cepstral Coefficients). In a work [5], a general approach to the classification of accents based on MFCC and neural networks is proposed. It involves the use of a sequence of transformations of the initial audio data.

In [2], it is proposed to perform additional data processing before extracting MFCC from the initial audio files to improve the prediction accuracy.

The authors also propose the use of a convolutional neural network to classify 3 languages. They compared the learning outcomes of their network with standard Gradient Boost and Random Forest algorithms. The result was an average prediction accuracy on the test set of 88% for the wrapping network, while both algorithms showed an average of 69.1% accuracy.

In this work, the approach proposed in [2] and [5] is used.

An Application for Recognizing English Accents Based on Deep Machine Learning

Defining Neural Network Requirements

There are more than 100 accents in English. The ideal option would be to use all the accents to learn the neural network. But this requires the necessary training data on these accents. On the other hand, such training requires huge computing and time resources.

Therefore, it w\as decided to choose the most common English accents. For example, Australia, the United States, Canada, and Ireland are among the top 10 countries for British immigration. It is also necessary to take into account England itself and the countries that are part of the United Kingdom, i. e. Scotland and Wales. Thus, the target accents for network training were formed:

- The Australian English accent,
- The American English accent,
- The NY English accent,
- The Canadian English accent,
- The Irish English accent,
- The posh English accent,
- The Welsh English accent,
- The Scottish English accent.

The neural network must determine its percentage accuracy to above mentioned list of language accents from the audio recording.

Defining Software Requirements

Using neural network, it was decided to automate the process of progress assessment in learning English or its certain accents, creating an appropriate web application.

Potential users of the software product have been identified — people whose progress in language learning needs to be assessed: emigrants, actors, teachers, students and others.

Basic requirements for web app:

- the ability for users to add their audio recordings to the system by adding existing files or directly through the user's voice recording;
- an opportunity for users to get an assessment of how close the audio recording is in relation to each of 8 accents;
- the ability for authorized users to interact with these audio recordings (listen, save locally or in the application, delete);
- the ability for authorized users to group audio recordings by goals and get its progress in pronunciation according to the selected goal.

It was also decided to implement TelegramBot for more convenient interaction with audio files, as Telegram is a very popular messenger, which has more than 500 million downloads, and also allows you to upload large files, up to 1,5 GB [6, 7].

Design and Implementation of a Neural Network and Software

It was decided to use the data processing pipeline using MFCC, proposed in [2] and [5] (Fig. 1).

Therefore, in order to design and implement a neural network and software the following steps are required:

1. choose a dataset for learning neural network;
2. preprocess the data;
3. design a neural network (methods);
4. implement the neural network and the corresponding software;
5. conduct training and testing of the network.

The features of each step are described below.

1. *Dataset*

Data was taken for analysis from the Internet resource English Multi-speaker Corpus for Voice

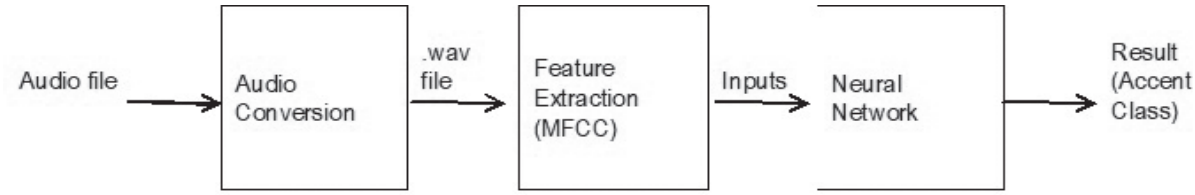

*Fig.*1. Approach using MFCC

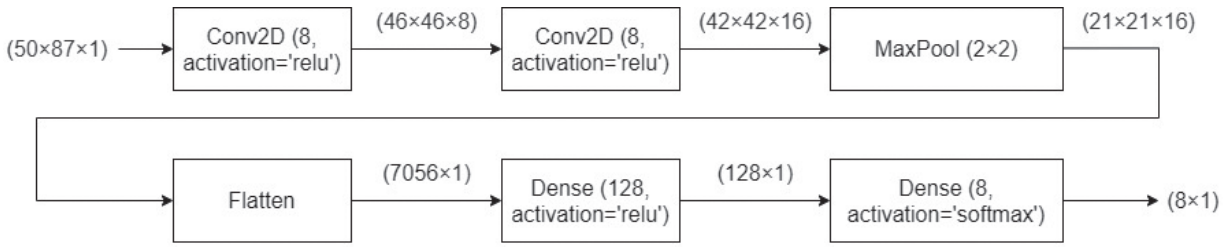

Fig. 2. Convolution network model

Cloning. Each speaker reads out about 400 different sentences, most of which were selected from a newspaper, where each set of sentences was selected using a greedy algorithm designed to maximize contextual and phonetic coverage.

Characteristics and advantages of the dataset:

- includes language data spoken by 109 native English speakers with different accents;
- contains a large number of different English accents, including the 8 accents that were selected in this work;
- includes the pronunciation of words that most fully convey the features of the sound of a particular accent;
- all speakers have exclusively English accents;
- includes the pronunciation of speakers of both sexes, which allows to teach the network to classify the pronunciation of any speaker and reduce the likelihood of obtaining an unpredictable assessment in the case of training on the pronunciation of speakers of only one sex.
- large data set — 10 GB;
- .wav audio file format.

Because the audio files in this dataset are in .wav format, one can skip the audio conversion step. In addition, the dataset contains demographic information about users, which indicates their accent.

2. *Preprocessing*

To ensure the accuracy of the future algorithm, it is necessary to provide speech data and get rid of "quiet" areas. To do this, a script was written, which, passing at short intervals, checks whether the energy density of this interval is higher than the average energy density of the entire audio signal multiplied by 4,8%. Energy density is calculated as the square sum of its amplitude on the current time window

$$E = \frac{A^2}{n \cdot m},$$

where A — matrix, that represents time window, $n \times m$ — its dimension.

After receiving exclusively speech from the recordings, it is necessary to perform the procedure of obtaining MFCC coefficients [8].

Each set of coefficients consists of 87 elements, i.e. the total dimension of the coefficient matrix obtained from 1 second of audio corresponds (50, 87).

The result of the procedure for obtaining MFCC coefficients in each audio recording is a tensor size (14542, 50, 87), i.e. 14542 results, each of 50 coefficients consisting of 87 values. The tensor was divided into train and test set in a ratio of 4:1, respectively.

3. *Methods*

The analysis of images traditionally makes use of convolutional neural networks [2], [5]. Since we have data in form of MFCC, we can feed them to the network for image analysis.

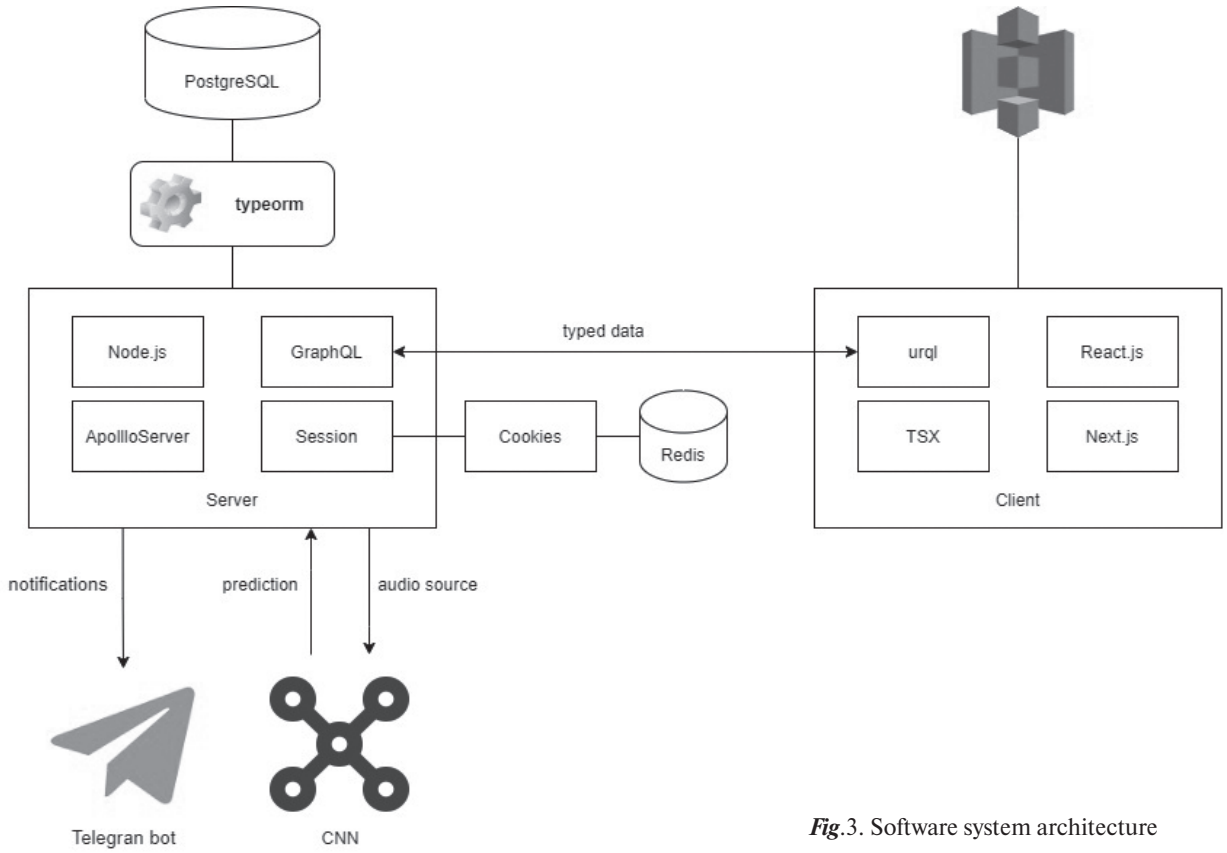

Fig.3. Software system architecture

The developed model of the neural network is presented in Fig.2.

Neural network characteristics:

- the network consists of 2 convolutional layers with 3x3 filters and ReLU activation functions, which adds nonlinearity to the network;
- the next layer is the max-pooling layer, which leaves the most information-rich neurons;
- The last layers are the so-called dense layers, which means that they are tightly connected to the source layer, to which the softmax activation function is applied. This function is used as the last in most modern artificial neural networks and performs the normalization of the results of the last layer of the network in the direct distribution of probabilities of each class (in our case, the accent).

4. *Software implementation*

The web application is built using client-server technology. In accordance with the above mentioned application requirements, the architecture contains following components (Fig. 3):

- CNN (Convolutional Neural Networks) — a neural network that evaluates accents;
- Telegram bot is written in Typescript language to simplify the work with audio recording and receiving notifications;
- PostgreSQL - a database for storing audio files data;
- the client part provides an interface for recording and listening to audio, creating goals and viewing their progress;
- the server part performs the functions of predicting accents, interaction with the database, processing user requests;
- AWS S3 — cloud storage for storing audio recorded by users;
- Redis is a distributed repository of key-value pairs that is used to cache data during a user session.

To implement the convolutional neural network, the Python programming language was used through easy and intuitive syntax and access to libraries for working with large amounts of data (pandas, numPy) and for processing audio files (librosa). The latter is often used to obtain MFCCs from .wav files [2], [5], [9].

Client part:

- implemented using React.js (framework for building a front-end part in Typescript);
- urql library was used to generate typed GraphQL queries to the server;
- The Next js library is used to facilitate work with page navigation;
- TSX — a special form of writing code in typescript, using HTML blocks as expressions, which is used to avoid creating static HTML pages.

Server part:

- implemented on Node.js (framework for building a back-end part in Typescript);
- GraphQL language is used to build a strictly typed API;
- ApolloServer is used to build a GraphQL server that allows you to operate on entities created by the developer, to perform CRUD operations on them.

5. *Results*

Model training took 80% of the total tensor. Other 20% was used to test the model across 2 epochs. Test data are presented in Table.

Table

Accent	Test Accuracy (%)
Australian English	76
American English	96
NY English	97
Canadian English	86
Irish English	94
Posh English	92
Welsh English	88
Scottish English	88

The worst accuracy is obtained with the Australian accent, presumably because this accent is very similar to the Irish.

While listening to audio data from the selected dataset, a small amount of completely blank audio was detected, where the speaker is silent. Measures have been taken to disregard such data in the training process, but even after that it has not been possible to weed out a small percentage of such data, which affects the accuracy of predictions.

Conclusions

Using Python and Typescript programming languages, GraphQL, urql and React technologies, a web application has been developed that recognizes the percentage of how close user audio belongs to each of 8 popular English accents on the basis of a built and trained convolutional neural network.

Possibilities for improving the implementation of the neural network and software system in the following versions have been identified:

- achieving higher prediction accuracy by adding more filters and increasing the training time of the neural network;
- reducing the training time of the neural network by adding a check that the audio recording belongs to the English language.

The architecture of the software system is flexible and can be extended by adding new modules that expand the functions of the system.

It is possible to increase the number of accents that the system evaluates, although this will require additional resources, as it is necessary to train the network from the beginning. At the same time there is a probability of deterioration of the forecast accuracy.

The materials of the article will be useful in solving problems of audio, video and graphic materials classification using neural networks.

REFERENCES

1. "The most spoken languages worldwide in 2021", *Statista*. [online] Available at: <https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/>. [Accessed 18 May 2021].
2. *Sheng L. M. A., Edmund M. W. X.,* 2017. Deep learning approach to accent classification, Stanford University Project Report. [online] Available at: <http://cs229.stanford.edu/proj2017/final-reports/5244230.pdf>. [Accessed 18 May 2021].
3. *Upadhyay R.*, 2017. Accent Classification Using Deep Belief Network, University of Mumbai. [Accessed 18 May 2021].
4. *Hernandez S., Bulitko V., Carleton S., Ensslin A., Goorimoorthee T.,* 2018. Deep learning for classification of speech accents in video games, AIIDE Workshops. [online] Available at: <http://ceur-ws.org/Vol-2282/EXAG_114.pdf>. [Accessed 18 May 2021].
5. *Duduka S., Jain H., Jain V., Prabhu H., Chawan P.,* 2020. "Accent Classification using Machine Learning", *International Research Journal of Engineering and Technology (IRJET)*, 07 (11), pp. 638–641. [online] Available at: <https://www.irjet.net/archives/V7/i11/IRJET-V7I11105.pdf>. [Accessed 18 May 2021].
6. "Telegram Revenue and Usage Statistics", *Business of Apps*, 2020. [online] Available at: <https://www.businessofapps.com/data/telegram-statistics/>. [Accessed 18 May 2021].
7. "Telegram vs WhatsApp: The Basics", *Uctoday*. [online] Available at: <https://www.uctoday.com/collaboration/telegram-vs-whatsapp-the-basics/>. [Accessed 18 May 2021].
8. "Mel Frequency Cepstral Coefficient (MFCC) tutorial", *Practical Cryptography*. [online] Available at: <http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/>. [Accessed 18 May 2021].
9. *Duduka S., Jain H., Jain V., Prabhu H., Chawan P.*, 2021. "A Neural Network Approach to Accent Classification", *International Research Journal of Engineering and Technology (IRJET)*, 08 (03), pp. 1175–1177. [online] Available at: <https://www.irjet.net/archives/V8/i3/IRJET-V8I3359.pdf>. [Accessed 18 May 2021].

Received 20.07.2021

ЛІТЕРАТУРА

1. The most spoken languages worldwide in 2021. Statista. URL: https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/. (Last accessed: 18.05.2021).
2. *Sheng L. M. A., Edmund M. W. X.* Deep learning approach to accent classification. Stanford University Project Report. 2017. URL: http://cs229.stanford.edu/proj2017/final-reports/5244230.pdf. (Last accessed: 18.05.2021).
3. *Upadhyay R.* Accent Classification Using Deep Belief Network. University of Mumbai, 2017. (Last accessed: 18.05.2021).
4. *Hernandez S., Bulitko V., Carleton S., Ensslin A., Goorimoorthee T.* Deep learning for classification of speech accents in video games. AIIDE Workshops, 2018. URL: http://ceur-ws.org/Vol-2282/EXAG_114.pdf. (Last accessed: 18.05.2021).
5. *Duduka S., Jain H., Jain V., Prabhu H., Chawan P.* Accent Classification using Machine Learning. International Research Journal of Engineering and Technology (IRJET). 07 (11). 2020. P. 638–641. URL: https://www.irjet.net/archives/V7/i11/IRJET-V7I11105.pdf. (Last accessed: 18.05.2021).
6. Telegram Revenue and Usage Statistics. Business of Apps. 2020. URL: https://www.businessofapps.com/data/telegram-statistics/. (Last accessed: 18.05.2021).
7. Telegram vs WhatsApp: The Basics. Uctoday. URL: https://www.uctoday.com/collaboration/telegram-vs-whatsapp-the-basics/. (Last accessed: 18.05.2021).
8. Mel Frequency Cepstral Coefficient (MFCC) tutorial. Practical Cryptography. URL: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/. (Last accessed: 18.05.2021).
9. *Duduka S., Jain H., Jain V., Prabhu H., Chawan P. A* Neural Network Approach to Accent Classification. International Research Journal of Engineering and Technology (IRJET). 08 (03). 2021. P. 1175–1177. URL: https://www.irjet.net/archives/V8/i3/IRJET-V8I3359.pdf. (Last accessed: 18.05.2021).

Надійшла 20.07.2021

А.В. Манохін, студент кафедри ПЗКС ФПМ НТУУ, Нац. техн. ун-т України
«Київський політехнічний інститут імені Ігоря Сікорського»,
03056, м. Київ, просп. Перемоги, 37, Україна,
manokhin.kpi@gmail.com
Н.А. Рибачок, кандидат технічних наук, старший викладач кафедри ПЗКС ФПМ НТУУ,
Нац. техн. ун-т України «Київський політехнічний інститут імені Ігоря Сікорського»,
03056, м. Київ, просп. Перемоги, 37, Україна,
rybachok@pzks.fpm.kpi.ua

РОЗПІЗНАВАННЯ АКЦЕНТІВ АНГЛІЙСЬКОЇ МОВИ З ВИКОРИСТАННЯМ ГЛИБИННОГО МАШИННОГО НАВЧАННЯ

Вступ. Розпізнавання акцентів користувачів є актуальною задачею як для покращення функціонування програмних систем, так і для людей, які вивчають певну мову чи її акценти.

Мета статті. Розроблення нейронної мережі та відповідного програмного забезпечення для розпізнавання восьми акцентів англійської мови.

Методи. Обрано датасет для навчання нейронної мережі. Здійснено початкову обробку даних, яка полягає у вилученні «тихих» ділянок. Спроектовано згорткову нейромережу, що складається із двох згорткових шарів, одного шару *max pooling*, а також двох щільних шарів. Нейромережа та відповідне ПЗ реалізовано програмно. Проведено тренування мережі протягом двох епох на множині 5 516 аудіозаписів, взятих із ресурсу *English Multi-speaker Corpus for Voice Cloning*.

Результати. Досягнуто точність прогнозування 89,07% на тестових даних, що представлялися 11 тис. матрицями *MFCC* розмірністю 50×87. Розроблено програмне забезпечення для визначення акцентів англійської мови, яке надає можливість користувачу через веб-інтерфейс або Телеграм-бот за рахунок використання загорткової нейромережі визначати відсоток належності аудіозапису до восьми найбільш розповсюджених англомовних акцентів.

Висновки. Матеріали статті будуть корисними при вирішенні задач класифікації аудіо-, відео- та графічних матеріалів із використанням нейронних мереж.

Ключові слова: *нейронна мережа, глибинне машинне навчання, розпізнавання акцентів, MFCC, CNN.*