UDC 004.934.8 **DOI:** 10.18664/ikszt.v29i1.300905

Vladyslav Kholiev, Olesia Barkovska

Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

MODEL OF THE SPEAKER IDENTIFICATION AND VERIFICATION SUBSYSTEM

Abstract. The paper is focused on the pressing **problem** of authentication and verification of speakers based on voice information, which plays an important role, for example, in online or remote communication and information exchange in all spheres of life, including scientific communication. The **aim** of this paper is to create a model of a speaker identification and verification subsystem. To achieve this goal, the following **tasks** were accomplished: the connection of the modules of the proposed model was explained, the voice information analysis module was explored, while ensuring the scalability of the system with a significant increase in the number of users, and the results were analyzed. The developed pseudo-ensemble-based neural network module was tested on a dataset prepared on the basis of the LibriSpeach corpus, an open English speech corpus based on the LirbiVox project of voluntarily provided audio books. The result of applying the developed module on the selected dataset is demonstrated, demonstrating that in order to implement the subsystem in a neural network training system, the proposed pseudo-ensemble should be trained on at least 120 epochs using noise reduction methods at the stage of audio sequence preprocessing.
Key words: machine learning; speaker diarization; classification; pseudo-ensemble; sincnet; librispeach; librivox.

Introduction

Paper [1] presented a system of scientific communication and knowledge sharing for young scientists. The system operates in several modes: data storage, data processing, and voice information access. These modes of operation are enabled by a number of subsystems.

The data processing mode operates on the basis of the following subsystems:

- subsystem of speakers' identification and verification
- subsystem for vectorization of scientific papers
- subsystem for clustering vectorized scientific papers by research topics using the method of textual proximity

The functional diagram of the proposed system of knowledge exchange of young scientists is shown in the figure 1.

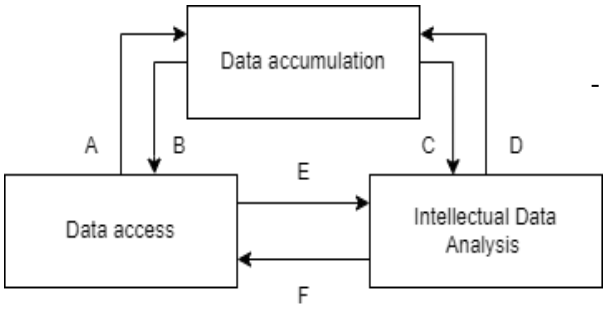


Figure 1 – Functional diagram of the proposed system of knowledge exchange of young scientists.

When considering the functionality of the system in each mode, it is important to understand what the inputs and outputs of each mode are:

marker A: data is accumulated in the form of user voice data for further training, as well as in the form of electronic documents of scientific papers for vectorization and further clustering of papers similar in topic.

marker B: the repository subsystem receives various requests: for access to scientific papers in various forms, for user data and metadata (for example, for the subsystem that provides social rooms), as well as requests for verification and identification of users.

marker C: the relevant modules receive voice data for training, or classification and research documents for further vectorization and clustering.

marker D: as a result of the analysis mode, neural network models are trained on the basis of voice data and their weights are saved to the repository, and based on the uploaded research documents, their vectorized representation is formed and saved, and the documents themselves are assigned to a group with similar topics.

marker E: if there is a need to identify or verify users, requests are sent with the relevant data (user ID(s) and/or conference IDs, etc.);

marker F: in response to the queries, either individual neural networks or a pseudo-ensemble module is deployed based on the stored weights of the trained networks.

This paper proposes a solution, the results of which can be used in subsystems that maintain virtual conference rooms and security, performing the following functions:

- diarization of a virtual conference by means of speaker identification.
- user authentication by verifying the speaker.

Пропозиція рішення описана у вигляді модуля на основі розподілених нейромережевих моделей, які використовують архітектуру SincNet [2].

An important aspect of modern life, present in an increasingly large part of it, is the authenticity of an individual and the ability to verify it. If earlier it was possible to uniquely identify a user by simple data such as email and password, nowadays, with computing power that is many times greater and allows to bypass more and more complex methods of personal identification, such data is not sufficient. For this reason, methods of identity identification and verification based on biometric data, such as fingerprints, retinas, face, and voice, are gaining popularity [3]. Among them, the voice identification method stands out due to its accessibility and convenience for the user.

Audio information plays one of the most widespread and important roles, as it is primarily used for communication and information exchange in all aspects of life, including scientific communication, or rather, inward-facing communication [4-5]. Scientific conferences, symposia, forums, etc. are mostly held in the format of audio reports or discussions accompanied by visual material, usually presentations. However, the accompanying visual material is mostly optional, and the participants' presentations are often designed to be listened to, so all the necessary information is contained in the report.

Diarization is an audio signal processing and transformation process that results in segmentation and labeling of the audio signal according to the sources of sound. In other words, diarization answers the question "who speaks when?" or "what sounds when?". This process generally consists of several stages, such as pre-processing the input audio to improve the quality and accuracy of the result, detecting and splitting the audio stream into segments when an audio event occurs, encoding (embedding) signal characteristics, and further clustering or classification depending on the task [6].

Over the years, algorithms and approaches to individual stages, or even to the process as a whole, have been changing and improving. Modern approaches to diarization include the so-called end-to-end neural diarization, where individual submodules of the traditional speaker diarization system mentioned above can be replaced by one or more neural networks [7-10]. With the rapid development of deep learning and ensemble approaches, this strategy is gaining popularity [11-13]. This paper focuses on the stages of audio classification and labeling.

Diarization is used to solve a variety of applications, such as speaker recognition, audio analysis and segmentation, and others (fig. 2):

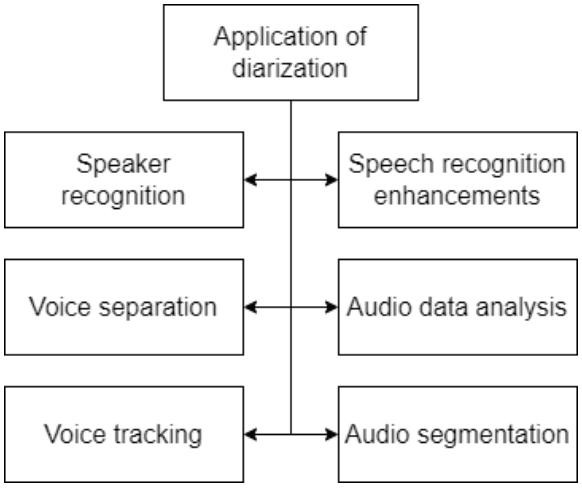


Figure 2 – Applications of diarization.

1. Speaker recognition: Identifying the voices of specific people in an audio recording.
2. Voice separation: Distinguishing between different voice sources in audio data.
3. Voice tracking: Tracking voice sources over time, useful in analyzing dynamic scenarios.
4. Speech recognition enhancement: Improving the accuracy of speech recognition systems by taking into account individual voice characteristics.
5. Audio data analysis: Identification of the time intervals in which speech occurs, as well as pauses and background sounds.
6. Audio segmentation: Dividing an audio recording into segments with homogeneous sound sources.

The aim of the work is to create a model of a speaker identification and verification subsystem. To achieve this goal, the following tasks must be done:

- To substantiate the interconnection of the modules of the proposed model;
- To investigate the module of voice information analysis, taking care to ensure the scalability of the system with a significant increase in the number of users;
- To analyze the obtained results

To accomplish these tasks, we propose a method of analyzing voice information based on artificial neural networks represented by a neural network pseudo-ensemble using the convolutional network SincNet. This method is a component of the voice information analysis module of the proposed model of the speaker identification and verification subsystem.

Materials and methods

The speaker identification and verification subsystem consists of the following modules:

- voice information analysis module;
- authentication module;

- module for interaction with social and conference rooms;

The paper proposes the analysis of voice information based on a distributed neural network pseudo-ensemble, which is a collection of individual, independent neural network models that receive the same input signal in waveform format in combination with a decision-making algorithm based on a voting algorithm.

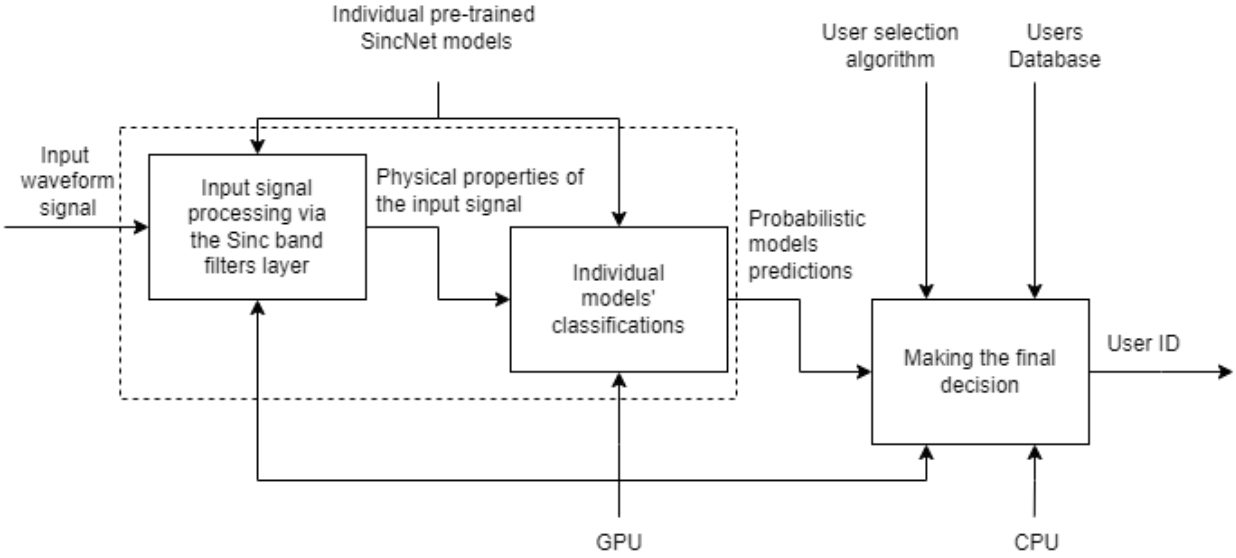


Figure 3 – Generalized block diagram of the speaker identification and verification module.

The function block (FB) processes the input signal through a layer of filters based on the mathematical filtering function sinc works on the basis of pre-trained SincNet networks [2]. SincNet is a CNN architecture in which the first layer is replaced by a convolution with sinc filters (fig. 4).

Unlike similar approaches, the SincNet network receives the signal in its original form, and bandpass filters are used as convolutions, the parameters of which are determined by the network during training. For each filter, only 2 parameters are trained - the high and low frequencies. Thus, the network learns data in the context of only certain frequency ranges, while working with the original signal, which reduces the necessary preprocessing.

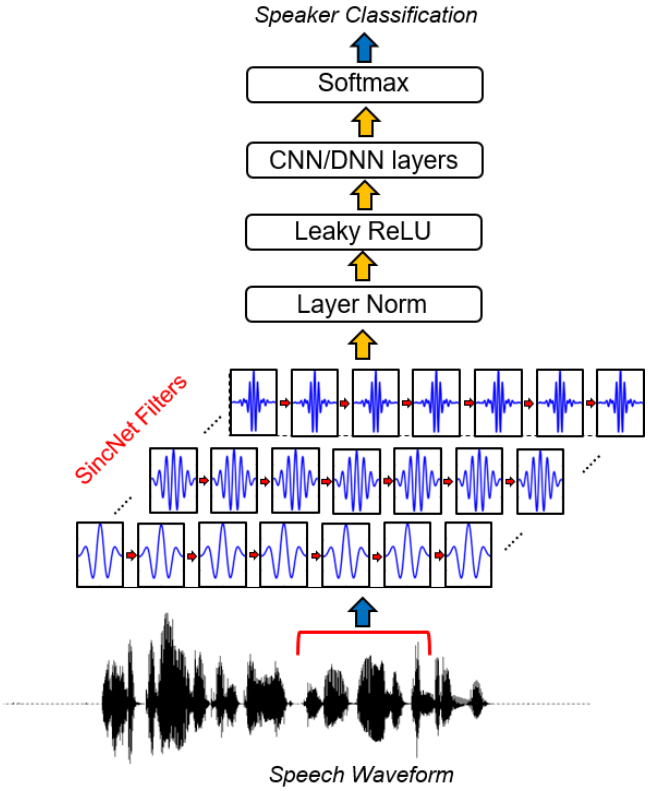


Figure 4 – SincNet model architecture [2, p.2]

Bandpass filters themselves are represented as the difference of two low-pass filters, and when translated into the time domain, the filter is the difference of two sine functions. Multiplying the original signal by the resulting convolution is equivalent to selecting a signal of a certain frequency band.

Thus, at the first level, SincNet learns filters with physical content based on the input signal in the audio waveform format. This approach provides SincNet with a number of advantages over conventional CNNs, such as faster convergence, fewer parameters, and less noisy filters (fig. 5).

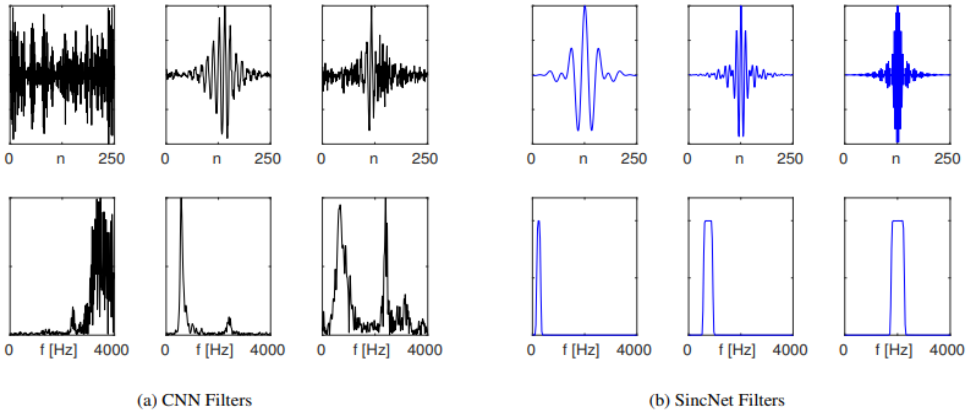


Figure 5 – Examples of filters learned by standard CNN and SincNet (using the Librispeech corpus). The first row shows the performance of the filters in the time domain, and the second line shows their amplitude-frequency response [2, p.4].

The final decision-making FB works as follows: each of the models in the system is fed the same audio stream as input, after which each of the models performs a binary classification (despite the basic capabilities of SincNet to perform multi-class classification). Each neural network trained to recognize a single speaker gives a numerical probability value of the speaker's belonging to the neural network class (user ID). After that, the probabilistic assumptions of individual models are collected into a collection and a decision is made at the third stage of the module (fig. 6). When interpreting the result, the weighted voting method is used, i.e., the class is determined by the number of the neural network that showed the maximum value at the output.

$$Id = max([p_1,\ p_2, \dots, p_n]), \quad (1)$$

where Id – internal identifier of the neural network, which corresponds to the internal identifier of the user,

p_x – probabilistic score of an individual neural network obtained as a result of classification,

$\overline{n[1,x]}$ – the number of neural networks (corresponding to the number of users) in the pseudo-ensemble at a particular point in time.

This algorithm is dictated by both the system concept and the intended generalized use case – only one model should identify the user, while the others should demonstrate low scores.

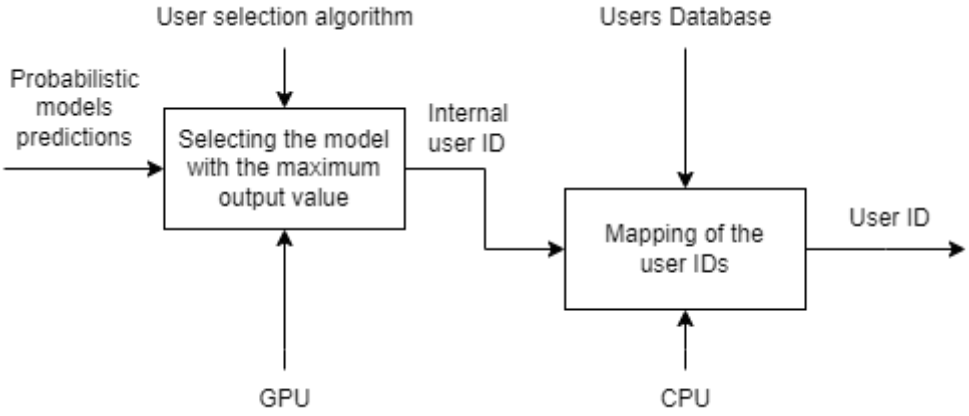


Figure 6 – Decision-making algorithm at the module level.

This approach provides flexibility and supports the idea of system modularity, as well as ensures that the system responds with a low degree of error in the event that more than one neural network has given a high probability of the input signal belonging to their user class (for example, if there are people with similar voices among users).

Experiments

As a dataset, we chose a dataset that was independently generated on the basis of the LibriSpeech

corpus [14], which, in turn, was generated on the basis of the LibriVox audiobook project [15], which is in the public domain, or rather its segments in English.

The corpus is divided into several parts, available separately: a test sample in a single archive, as well as a training sample in three archives – 100-, 300-, and 500-hour archives. In total, the corpus includes 982 hours from 2338 speakers.

Table 1. – *Data subsets in LibriSpeech*

subset	hou	per-speake minutes	female speaker	male speake	total speake
dev-clean	5.4	8	20	20	40
test-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-other	5.1	10	17	16	33
train-clean-100	100.	25	125	126	251
train-clean-360	363.	25	439	482	921
train-other-500	496.	30	564	602	1166

The authors of the corpus have created several annotation files: matching speakers and their characteristics with the subsets to which they are assigned, matching speakers and chapters of books whose records were taken for the corpus, and matching the books themselves with their identifiers used in other annotation files.

From these parts, subsets of 100 and 300 hours were taken, totaling 464.2 hours and 1172 speakers, which is approximately half of the total volume.

Initially, the corpus was created for the task of speech recognition, as well as identification and/or classification of certain characteristics of speakers (age, gender, etc.). To create the corpus, two stages of alignment were first performed using a variety of tools and speech models [9], which were aimed at dividing the recordings into fragments and removing recordings that contained discrepancies with the text due to human error (inclusions, substitutions, deletions, and permutations). These alignment steps resulted in approximately 1200 hours of recordings up to 35 seconds long, after which the data was segmented into smaller fragments based on pauses of at least 0.3 seconds. The test data was segmented similarly to the training data, but with the additional condition of splitting fragments only at sentence boundaries to better model language usage.

Due to additional pre- and post-processing stages, the balance of speakers' genders was ensured, and cases of recordings with multiple speakers were eliminated (by filtering individual recordings as well as certain genres that by their nature land themselves to multi-speaker recordings).

```
ID   |SEX| SUBSET            |MINUTES| NAME
14   | F | train-clean-360   | 25.03 | Kristin LeMoine
16   | F | train-clean-360   | 25.11 | Alys AtteWater
17   | M | train-clean-360   | 25.04 | Gord Mackenzie
20   | F | train-other-500   | 30.07 | Gesine
23   | F | train-clean-360   | 25.23 | Anita Roy Dobbs
25   | M | train-other-500   | 30.16 | John Gonzalez
26   | M | train-clean-100   | 25.08 | Denny Sayers
27   | M | train-clean-100   | 20.14 | Sean McKinley
28   | F | train-clean-360   | 25.03 | Kristin Hughes
29   | M | train-other-500   | 30.10 | Linton
31   | M | train-other-500   | 23.79 | Martin Clifton
32   | F | train-clean-100   | 24.01 | Betsie Bush
36   | M | train-other-500   | 25.85 | Chip
...
8824 | M | train-clean-360   | 25.21 | Mark Johnston
8825 | F | train-clean-360   | 23.93 | Erin Schellhase
8838 | M | train-clean-100   | 25.06 | Kevin Owens
8855 | M | train-clean-360   | 25.01 | Eric Metzler
8975 | F | train-clean-100   | 25.11 | Daisy Flaim
9022 | F | train-clean-360   | 25.17 | Claire M
9023 | F | train-clean-360   | 25.19 | P. J. Morgan
9026 | F | train-clean-360   | 21.75 | Tammy Porter
```

Listing. 1. – A fragment of one of the annotation files

As mentioned in Section 3.3, each of the models in the system is fed the same audio stream. This is done in the same way as in [2], i.e., each audio recording with a speaker is divided into short fragments (up to 200 ms) to

ensure that only one speaker is present in one audio fragment, and fed to the inputs of individual models. Other parameters and initialization values are also used similar to those presented in the original work, as they are good baseline values, namely: 80 filters of length 251 were used for the band pass filter layer, with the Xavier initialization method.

As a result, a vector of predictions from individual models is obtained for each fragment and the class with the highest probability is selected. Based on the classes, most fragments are assigned a speaker class for recording. For example, if the output of the module is a vector of output values [0.2, 0.6, 0.4], the second component of the vector has the maximum value. So, the class to which this example belongs will be 2.

Results

The results of the tests are provided in Table 2.

Table 2. – *Багатофакторна матриця планування тестів*

Test	Number speake	Number of	Levels of addi noise, dE	Average accur
1	2	60	0	94.02
2			25	92.13
3			50	74.14
4		80	0	95.16
5			25	93.67
6			50	91.17
7		120	0	**98.70**
8			25	**96.36**
9			50	**93.60**
10	5	60	0	93.45
11			25	92.07
12			50	71.34
13		80	0	94.80
14			25	92.15
15			50	89.92
16		120	0	**98.55**
17			25	**96.03**
18			50	**93.77**
19	10	60	0	93.33
20			25	91.21
21			50	71.04
22		80	0	94.32
23			25	92.58
24			50	89.27
25		120	0	**98.47**
26			25	**96.63**
27			50	**93.89**

As shown in the multifactorial planning table, 27 tests were conducted, which differed in the following factors:

- number of speakers (varies from two to ten);
- number of epochs (60, 80, 120) during which the neural networks were trained;
- three levels of additional noise (no noise, 25dB, and 50dB).

The accuracy of the entire pseudo-ensemble is calculated using the arithmetic mean formula:

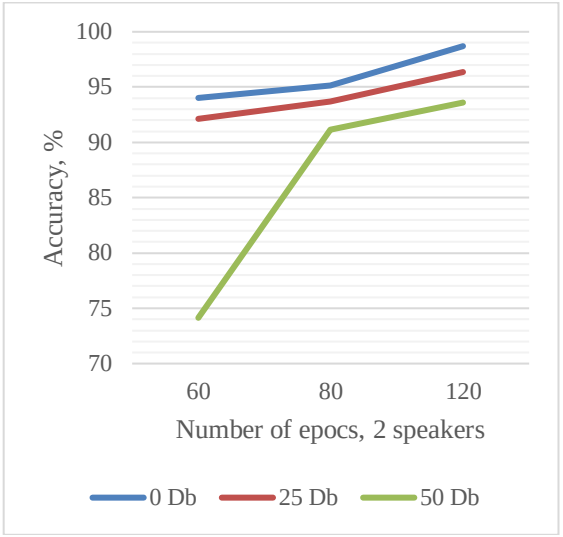
$$Acc_{avg} = \frac{1}{K}\sum_{i=1}^{K} Acc_i, \quad (2)$$

where Acc_{avg} is arithmetic mean of the accuracy of individual neural networks,

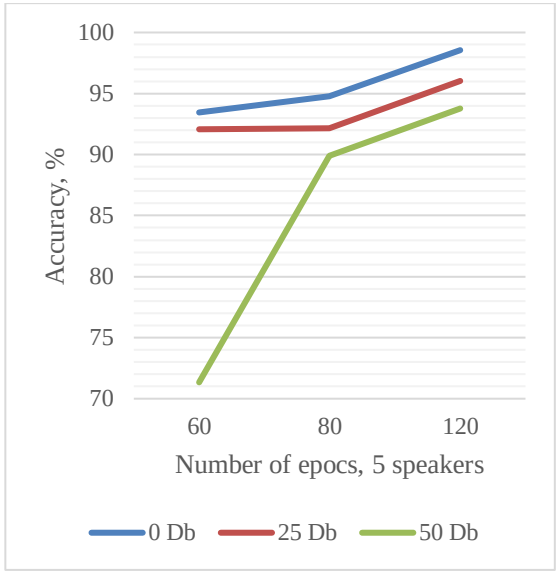
Acc_x – accuracy index of an individual neural network,

$\overline{K[1,x]}$ – the number of neural networks (corresponding to the number of users) in the pseudo-ensemble at a particular time.

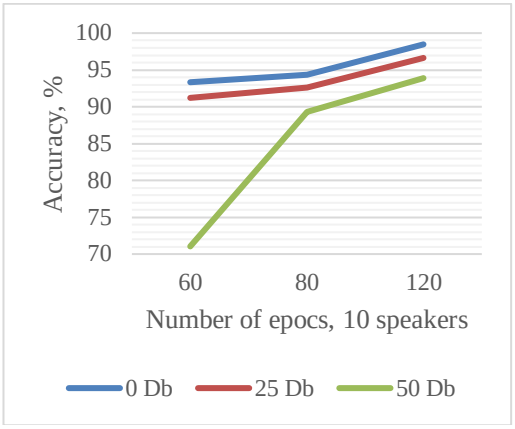
It is apparent that there is a correlation between accuracy, noise level, number of epochs, and number of speakers, especially in the effect of noise level, which suggests the need to pay more attention to noise reduction and isolation of the speaker's audio.



(а) 2 speakers



(б) 5 speakers



(в) 10 speakers

Figure 7 – Graphs of accuracy versus noise level for different number of epochs

With an increase in epochs, the accuracy rate increases by an average of 10.3%, which is logically consistent with the principles of deep learning, according to which longer training time leads to better learning of features, which in turn results in a better accuracy rate.

At the same time, as the level of speakers increases, the accuracy slightly decreases by 0.91% on average, as more networks lead to more errors.

The effect of noise level on accuracy is negative. As the noise level increases, the accuracy drops by 10.2% on average, regardless of the number of epochs and speakers.

Conclusions

In this paper, a model of the voice information analysis module of the speaker identification and verification subsystem was proposed, which will be used in the virtual conference room and security subsystems to perform diarization and authentication functions. This will ensure the scalability of the subsystem with an increased number of users, thanks to the proposed method of analyzing voice information based on artificial neural networks, represented by a neural network pseudo-ensemble using the convolutional network SincNet, which is capable of replacing individual neural network models in its composition, which in turn eliminates the need to re-train a single neural network model for each new user (assuming that the volume of users to be added is calculated in hundreds), as well as greater reliability and robustness.

The obtained results show the negative impact of noise on the quality of speaker classification (on average by 10.2% when the noise increases from 0dB to 50dB), which is partially compensated for by increasing the training epochs of the neural networks that make up the proposed pseudo-ensemble (on average by 10.3% when the number of training epochs increases from 60 to 120). This allows us to conclude that in order to implement the subsystem in the system, training of neural networks as part of the proposed pseudo-ensemble should be performed at least on 120 epochs, using noise reduction methods at the stage of audio pre-processing.

Since the main focus of the work was on the classification stage itself, future improvements will be focused on the decision-making unit, and more specifically on the ability to support simultaneous speech by multiple speakers, as well as combining the proposed module with audio signal pre-processing approaches to ensure the best performance of the module.

References

1. Холєв В., Барковська О. COMPARATIVE ANALYSIS OF NEURAL NETWORK MODELS FOR THE PROBLEM OF SPEAKER RECOGNITION //СУЧАСНИЙ СТАН НАУКОВИХ ДОСЛІДЖЕНЬ ТА ТЕХНОЛОГІЙ В ПРОМИСЛОВОСТІ. – 2023. – №. 2 (24). – С. 172-178.
2. Ravanelli, M., Bengio, Y. (2018), "Speaker Recognition from Raw Waveform with SincNet", 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, P. 1021–1028. DOI: https://doi.org/10.1109/SLT.2018.8639585
3. Olesia, B., Iana, M., Nataliia, Y., Oleksii, L., & Danyil, T. (2019). System of individual multidimensional biometric authentication. International Journal of Emerging Trends in Engineering Research, 7(12), 812-817.
4. Illingworth, S.; Allen, G. (2020), "Introduction", Effective science communication: a practical guide to surviving as a scientist (2nd ed.), Bristol, UK; Philadelphia: IOP Publishing. P. 1–5. DOI: https://doi.org/10.1088/978-0-7503-2520-2ch1
5. Côté, I., Darling, E. (2018), "Scientists on Twitter: Preaching to the choir or singing from the rooftops?", FACETS, 3. P. 682–694. DOI: https://doi.org/10.1139/facets-2018-0002
6. Mane, A., Bhopale, J., Motghare, R., & Chimurkar, P. An Overview of Speaker Recognition and Implementation of Speaker Diarization with Transcription. International Journal of Computer Applications, 975, 8887.
7. Kahn, J., Lee, A., & Hannun, A. (2020, May). Self-training for end-to-end speech recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7084-7088). IEEE.
8. Fujita, Y., Kanda, N., Horiguchi, S., Xue, Y., Nagamatsu, K., & Watanabe, S. (2019, December). End-to-end neural speaker diarization with self-attention. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 296-303). IEEE.
9. Fujita, Y., Kanda, N., Horiguchi, S., Nagamatsu, K., & Watanabe, S. (2019). End-to-end neural speaker diarization with permutation-free objectives. arXiv preprint arXiv:1909.05952.
10. Horiguchi, S., Fujita, Y., Watanabe, S., Xue, Y., & Nagamatsu, K. (2020). End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors. arXiv preprint arXiv:2005.09921.

11. Park T. J. et al. A review of speaker diarization: Recent advances with deep learning //Computer Speech & Language. – 2022. – Т. 72. – С. 101317.
12. Dhanjal, A. S., & Singh, W. (2023). A comprehensive survey on automatic speech recognition using neural networks. Multimedia Tools and Applications, 1-46.
13. Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., ... & Almojil, M. (2021). Automatic speech recognition: Systematic literature review. IEEE Access, 9, 131858-131876.
14. Kholiev, V., Barkovska, O. (2023), "Analysis of the of training and test data distribution for audio series classification", Information and control systems at railway transport, No. 1, P. 38-43. DOI: https://doi.org/10.18664/ikszt.v28i1.276343
15. V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.

Received (Надійшла) 15.12.2023.

Модель підсистеми ідентифікації та верифікації спікерів

Vladyslav Kholiev, Olesia Barkovska

Анотація. Робота присвячена актуальній **проблемі** автентифікації та верифікації спікерів за голосовою інформацією, що відіграє важливу роль, наприклад, при онлайн чи дистанційному спілкуванні та обміні інформацією в усіх сферах життя, включаючи наукову комунікацію. **Метою** даної роботи є створення моделі підсистеми ідентифікації та верифікації спікерів. Для досягнення цієї мети були виконані наступні **задачі**: обґрунтовано зв'язок модулів запропонованої моделі, досліджено модуль аналізу голосової інформації, з дотриманням забезпечення масштабування системи при значному збільшенні кількості користувачів та проаналізовано отримані результати. Розроблений нейромережевий модуль на основі псевдо-ансамблю був протестований на наборі даних, що був підготовлений на основі корпусу LibriSpeach – відкритого корпусу англійського мовлення, заснованого на проекті добровільно наданих аудіокниг LirbiVox. Продемонстрований результат застосування розробленого модуля на обраному наборі даних, який показує що для імплементації підсистеми в систему навчання нейронних мереж у складі запропонованого псевдо ансамблю повинно виконуватися не менш ніж на 120 епохах, використовуючи методи шумопригнічення на етапі попередньої обробки аудіоряду.

Ключові слова: machine learning; speaker diarization; classification; pseudo-ensemble; sincnet; librispeach; librivox

ABOUT THE AUTHORS / ВІДОМОСТІ ПРО АВТОРІВ

Холєв Владислав Олександрович – асистент кафедри "Електронно обчислювальних машин", Національний університет радіоелектроніки «ХНУРЕ», Харків, Україна;

Барковська Олеся Ю'ївна – кандидатка технічних наук, доцентка кафедри "Електронно обчислювальних машин", Національний університет радіоелектроніки «ХНУРЕ», Харків, Україна;

Vladyslav Kholiev – Professor Assistant at Electronic Computers Department, Kharkiv National University of Radio Electronics "NURE", Kharkiv, Ukraine.
e-mail: vladyslav.kholiev@nure.ua; ORCID ID: https://orcid.org/0000-0002-9148-1561.

Olesia Barkovska - Assoc. Professor at Electronic Computers Department, Kharkiv National University of Radio Electronics "NURE", Kharkiv, Ukraine.
e-mail: olesia.barkovska@nure.ua; ORCID ID: https://orcid.org/0000-0001-7496-4353