ISSN 2522-9818 (print)
ISSN 2524-2296 (online)
Innovative technologies and scientific solutions for industries. 2023. No. 2 (24)

UDC 004.934.8
DOI: https://doi.org/10.30837/ITSSI.2023.24.172

V. KHOLIEV, O. BARKOVSKA

COMPARATIVE ANALYSIS OF NEURAL NETWORK MODELS FOR THE PROBLEM OF SPEAKER RECOGNITION

The **subject matter** of the article are the neural network models designed or adapted for the problem of voice analysis in the context of the speaker identification and verification tasks. The **goal** of this work is to perform a comparative analysis of relevant neural network models in order to determine the model(s) that best meet the chosen formulated criteria, – model type, programming language of model's implementation, parallelizing potential, binary or multiclass, accuracy and computing complexity. Some of these criteria were chosen because of universal importance, regardless of particular application, such as accuracy and computational complexity. Others were chosen due to the architecture and challenges of the scientific communication system mentioned in the work that performs tasks of the speaker identification and verification. The **relevance** of the paper lies in the prevalence of audio as a communication medium, which results in a wide range of practical applications of audio intelligence in various fields of human activity (business, law, military), as well as in the necessity of enabling and encouraging efficient environment for inward-facing audio-based scientific communication among young scientists in order for them to accelerate their research and to acquire scientific communication skills. To achieve the goal, the following **tasks** were solved: criteria for models to be judged upon were formulated based on the needs and challenges of the proposed model; the models, designed for the problems of speaker identification and verification, according to formulated criteria were reviewed with the **results** compiled into a comprehensive table; optimal models were determined in accordance with the formulated criteria. The following neural network based **models** have been reviewed: SincNet, VGGVox, Jasper, TitaNet, SpeakerNet, ECAPA_TDNN. **Conclusions**. For the future research and practical solution of the problem of speaker authentication it will be reasonable to use a convolutional neural network implemented in the Python programming language, as it offers a wide variety of development tools and libraries to utilize.

Keywords: comparative analysis; neural network; intellectual models; model; machine learning; speaker identification; speaker recognition.

Introduction

Despite the rapid spread of the Internet at the beginning of the 21st century and the predominantly textual nature of the information that circulated on it at the beginning of its development, a significant part of the information generated, transmitted, and consumed by humanity remained audiovisual in nature. This is due not only to the limitations of the Internet technology at the time, but also to the biological characteristics of humans as a species, since most of the information we received from the environment is visual and sound information.

Over time, this trend has not only persisted, but also deepened with the development of technologies for generating, transmitting, and storing information. In turn, information processing and analysis technologies have developed and continue to develop still. The degree of decision-making automation continues to grow with the use of deep learning technologies and statistical models.

Recognition and processing of signals such as image, video, audio, and text are highly relevant areas of research that share the need to process large amounts of data, often followed by the application of intelligent analysis techniques [1–3].

As mentioned above, audio information in particular plays one of the most widespread and important roles as it is primarily used in communication and information sharing in all spheres of life, including scientific communication, more specifically inward-facing communication [4–5]. Scientific conferences, symposiums, forums etc. are held predominantly in the format of audio reports or discussions, accompanied by visual material, usually presentation. However, the accompanying visual material is largely optional, and members' presentations are often designed to be listened to with all the necessary information included in the report.

The tasks associated with processing and analyzing audio information include the following:

- voice cloning [6];
- speaker (voice) recognition [7];
- speaker emotion recognition [8].
- speech analytics (search for keywords in a dialog, call assessment, smart assistant, speech annotation, subtitling, speech recognition, transcription of lectures and meetings, etc;)
- etc.

Reviewing the popular current tasks, two similar tasks can be seen – speaker recognition and speech recognition. The difference lies in the fact that during speech recognition, the stage of transcription of speech into text is important, while speaker recognition focuses on analyzing the unique characteristics of speech due to human anatomy. Both technologies first appeared about 60 years ago but became available and effective only with the development of machine learning and the growth of computing power and cloud technologies.

With this information in mind, it becomes clear that enabling and encouraging efficient environment for inward-facing audio-based scientific communication among young scientists is important for them to accelerate research and acquiring scientific communication skills.

Analysis of last achievements and publications

For the matter of scientific communication, there doesn't exist a platform that combines the functionality of paper sharing database as well as a communication environment where two aspects are integrated into each other. With the abundance of online communication solutions, the task of audio communication has been relegated to various external products (e.g. Zoom, Google Meet etc.). This approach, however, is heavily catered towards the experienced researchers who possess the skills and know-how to navigate the vast network of diverse scientific communities and events that offer an opportunity to communicate via publication, presentation and discussion of their work and finding. Meanwhile, young scientists without aforementioned assets are rendered to work in the environment where they are told that they must 'publish or perish', and training in efficient communication often relies on a baptism of fire [4].

Such an International System of Knowledge Exchange of young scientists (ISKE) is proposed in [9]. As shown on figure 1, the system operates in several modes (data collection and processing mode, access mode) and consists of modules that perform different functions, one of which is a subsystem of social rooms for young scientists grouped by another module by scientific interests and research areas with the possibility of holding conferences. These conferences involve transcription with speaker division and subsequent annotation. This subsystem is also used to verify users when they log in, i.e. as an additional layer of authentication.

System analyses members' papers and qualification works in order to group them by areas of research and interests using methods for determining vectorized text proximity. This enables organization of a community of people with certain scientific interests that in turn encourages their further communication with their peers and potential integration into international research projects.

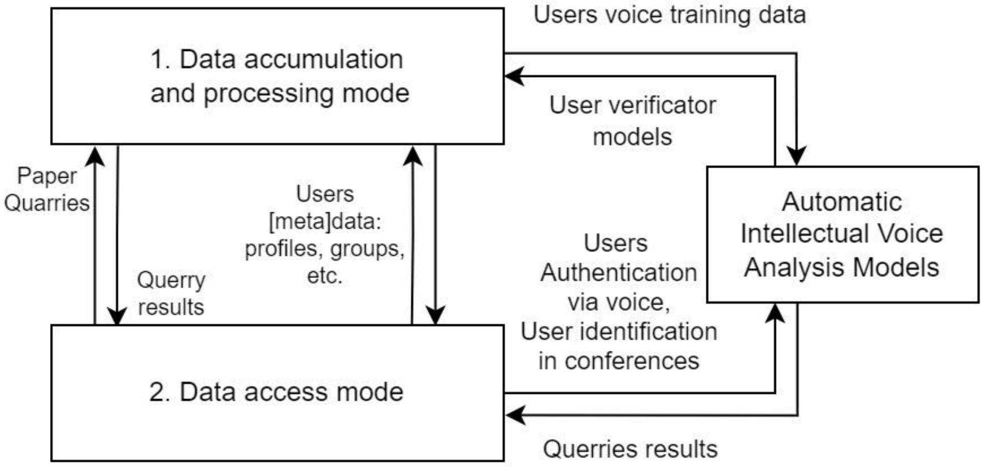
1. Data accumulation and processing mode
Users voice training data
User verificator models
Automatic Intellectual Voice Analysis Models
Paper Quarries
Querry results
Users [meta]data: profiles, groups, etc.
Users Authentication via voice, User identification in conferences
2. Data access mode
Querries results

Fig. 1. The system of knowledge exchange of young scientists

Both speaker verification and speaker identification belong to a set of problems known as voice recognition. There are different techniques that attempt to solve this set of problems that can be broadly divided into automatic and manual (Fig. 2). The latter, including for example stenography, – a process of writing in an abbreviated symbolic writing method, called shorthand, – were used before the widespread of information technologies, especially recent automatic techniques, which can be divided into neural network based and non-neural network based (for example, Support vector machines and Random forest).

The paper attempts to conduct a comparative analysis of automatic neural network based techniques, that manifest in various models.

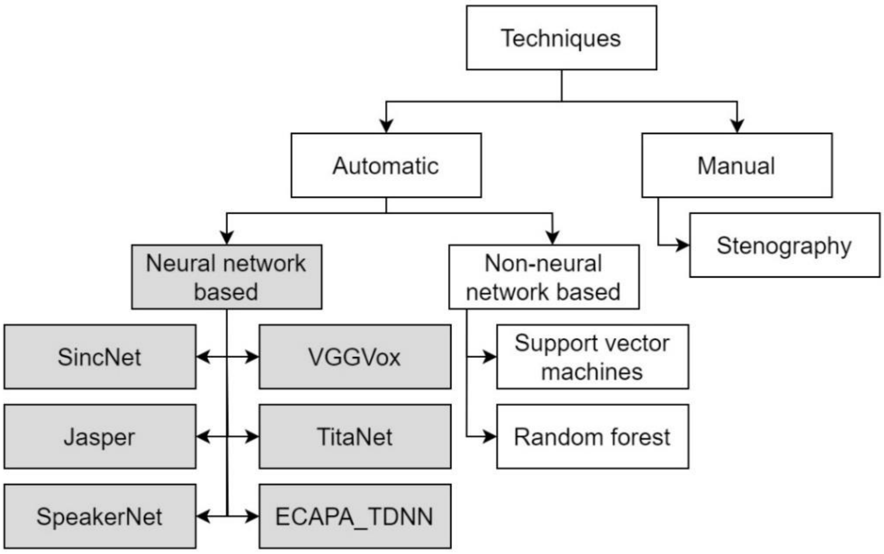

Fig. 2. Classification of voice recognition techniques

The aim of the work is to perform the comparative analysis of neural network models in the context of speaker identification and verification problem according to selected criteria. To achieve this goal, the following tasks are to be accomplished:

– to formulate criteria for models to be judged upon;

– to review the models, designed for the problems of speaker identification and verification, according to formulated criteria;

– to determine corresponding model(s) in accordance with the formulated criteria of the needed to be solved task.

Therefore, the task of comparative analysis of neural network models in the context of speaker identification and verification problem is a relevant task, due to importance of establishing an efficient and encouraging environment for inwards-facing scientific communication for young scientist.

Further development includes adapting the aforementioned models to the task of speech separation with the consideration of their corresponding results, that will be achieved in the summary table (Table 1).

Materials and methods

Among the tasks that the system presented in [6] needs to perform utilizing its modules, such as text vectorization and clustering based on topic proximity, automatic annotation of papers and held conferences, data security, etc. this paper puts focuses on the tasks of speaker recognition and identification. To solve these tasks neural network based techniques should be utilized, that best suit certain criteria. Such criteria are necessary to work out so that they can correspond to the SIMD execution model due to the system's features and challenges related to ensuring the efficient operation of its modules.

These requirements can be satisfied by meeting the following criteria:

– model type (CNN, RNN, etc.) or architecture it’s based on. The type of the model often informs us about broad and general advantages, drawbacks as well as typical applications of the model;

– programming language of model’s implementation. Self-explanatory;

– parallelizing potential. Due to the proposed models’ other tasks’ needs as well as general benefits of parallel computation, this critetion goes hand in hand with the previous one, beause of varying availability of development tools acroos different programming languages;

– whether the model is binary or multiclass as it will affect it’s use cases within the communication system;

– accuracy. Self-explanatory;

– computing complexity. Not only the system consists of many modules each with their own task that require the computation capacity of the hardware that the system will be deployed to which limits the, but the tasks of speaker recognition are time sensitive, which greatly limits

Below we analyze the neural network models according to the proposed criteria.

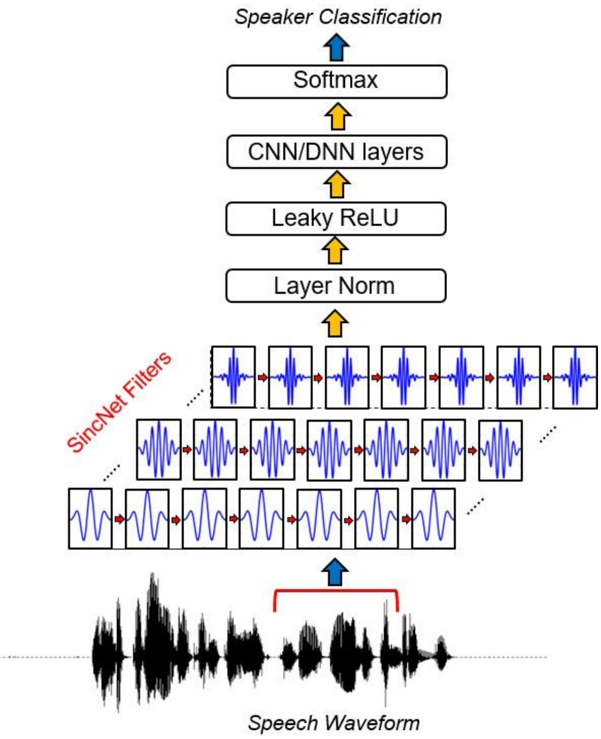

Fig. 3. SincNet model architecture [10, p. 2]

SincNet

The SincNet model is a model derived from Convolutional neural network (CNN). It is based on parametrized sinc functions, which implement band-pass filters. In contrast to standard CNNs, t hat learn all elements of each filter, only low and high cutoff frequencies are directly learned from data with such method. This offers a way to derive a customized filter bank specifically tuned for the desired application [10]. The architecture of the model is presented in figure 3.

VGGVox

The VGGVox model is a deep CNN based neural speaker embedding system, trained to map voice spectrograms to a compact Euclidean space where distances directly correspond to a measure of speaker similarity. Once such a space has been produced, other tasks such as speaker verification, clustering and diarization can be straightforwardly implemented using standard techniques, with embeddings proposed in the paper as features [11–12].

TitaNet

The TitaNet model is based on the ContextNet architecture for extracting speaker representations [13]. It functions by employing 1D depth-wise separable convolutions with Squeeze-and-Excitation (SE) layers with global context followed by channel attention-based statistics pooling layer to map variable-length utterances to a fixed-length embedding (tvector). TitaNet is a scalable architecture designed primarily for speaker verification and diarization tasks. The structure of the model is presented in figure 4.

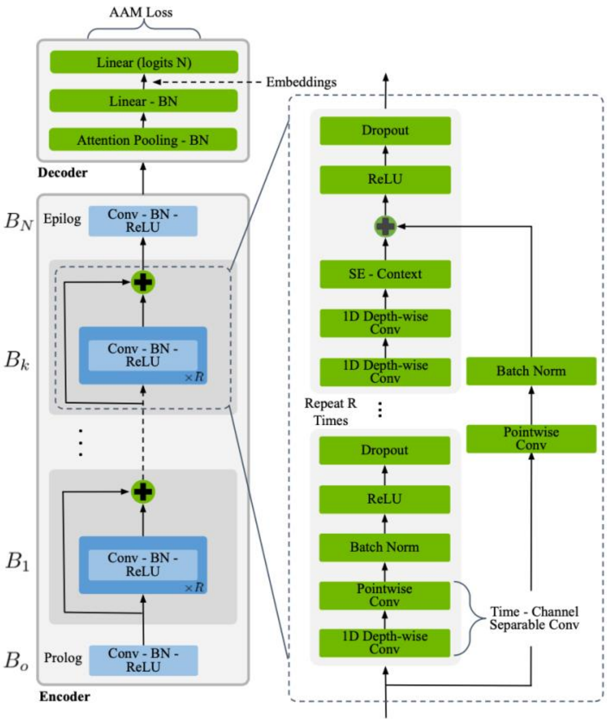

Fig. 4. TitaNet model structure [13, p. 2]

ECAPA_TDNN

The ECAPA-TDNN model is comprised of an encoder of time dilation layers that are based on Emphasized Channel Attention, Propagation, and Aggregation. It employs a channel and context dependent attention mechanism, Multi-layer Feature Aggregation (MFA), as well as Squeeze-Excitation (SE) and residual blocks [15]. Block diagram of the ECAPA-TDNN model is shown in figure 5.

Jasper

Jasper is an End-to-End Convolutional Neural Acoustic Model. It consists of several jasper blocks with 1 pre and 3 post conv layers. A Jasper BxR model has B blocks, each with R subblocks. Each sub-block applies the following operations: a 1Dconv, batch norm, ReLU, and dropout. All sub-blocks in a block have the same number of output channels. Each block input is connected directly into the last subblock via a residual connection. The residual connection is first projected through a 1x1 convolution to account for different numbers of input and output channels, then through a batch norm layer. The output of this batch norm layer is added to the output of the batch norm layer in the last sub-block. The result of this sum is passed through the activation function and dropout to produce the output of the current block [16]. The model architecture is shown in figure 6.

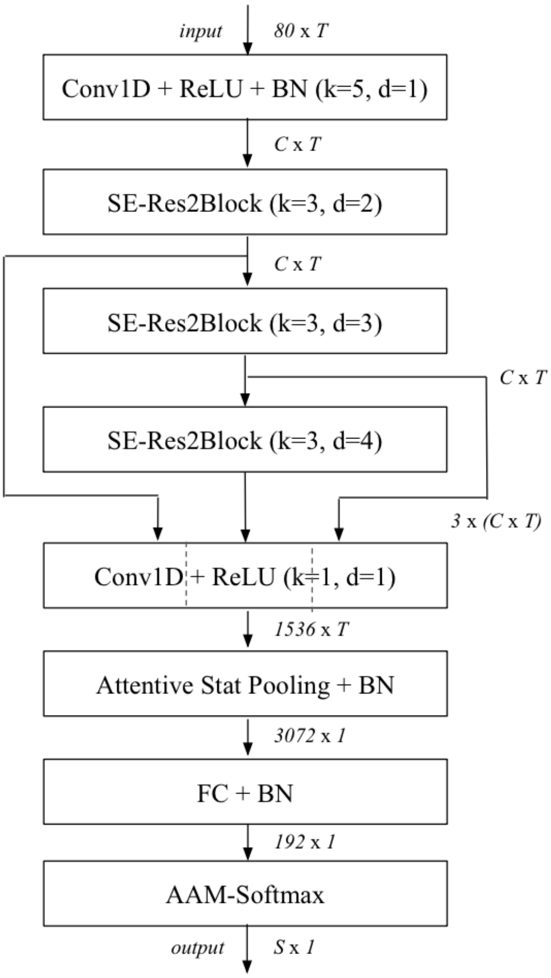

Fig. 5. Block diagram of the ECAPA-TDNN model [15, p. 3561]

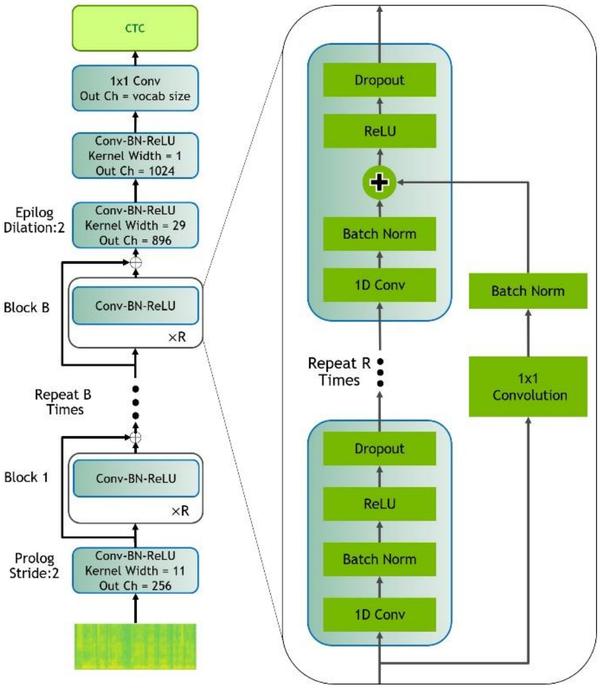

Fig. 6. Jasper model architecture [16, p. 72]

The results of the analysis are summarized in Table 1.

Table 1. *Results of comparative analysis against selected criteria*

Model name	Programming Language	Model Type	Binary or multiclass	Parallelization potential	Accuracy (Error rate)	Computational complexity
SincNet	Python	CNN	Both	Medium	0.32–0.96	Low
VGGVox	Python	ResNet	Multiclass	Medium	3.93–7.35	Low
Jasper	Python	CNN	Both	High	2.98–3.46	Medium to High
TitaNet	Python	CNN	Both	High	0.68–2.97	Medium to High
SpeakerNet	Python	CNN	Both	Hight	1.22–2.29	Medium
ECAPA_TDNN	Python	TDNN	Multiclass	High	2.64–3.66	Medium

Conclusion

In this paper, the relevance of the intelligent audio signal analysis problem was demonstrated, which manifests itself in tasks such as speech recognition and speaker recognition due to the widespread use of audio as a communication medium and in a wide range of practical applications in various fields of human activity.

The following criteria were proposed to evaluate existing neural network models for solving these tasks: the type of model or architecture on which it is based, the programming language of the model implementation, the parallelization potential, whether the model is binary or multiclass, accuracy, and computational complexity. These criteria were chosen due to the architecture and challenges of the scientific communication system mentioned in the work. It performs tasks of the speaker identification and verification.

Based on goal of the paper (to determine the model(s) to utilize in the scientific communication platform for the purpose of solving the problem of speaker recognition) and the obtained results, shown in Table 1, a conclusion can be made that for future research and practical solution of the problem it will be reasonable to use the SuncNet or convolutional neural network based on it implemented in the Python programming language, as it offers a wide variety of development tools and libraries to utilize, including CUDA architecture which enables efficient parallelization. This choice is motivated by its low computational complexity, potential for parallelism, which is moderate, and an accuracy metric that outperforms the second most accurate model analyzed (TitaNet) by at least 0.3%.

Further development includes adapting the reviewed models to the task of speech separation to be used in one of system's modules targeted at conferences with the consideration of the obtained results.

References

1. Barkovska, O. (2022), "Research into speech-to-text tranfromation module in the proposed model of a speaker's automatic speech annotation", *Innovative Technologies and Scientific Solutions for Industries*, No. 4 (22), P. 5–13. DOI: https://doi.org/10.30837/ITSSI.2022.22.005

2. Yashina, E., Artiukh, R., Pan, N., Zelensky, A. (2019), "Information technology for recognition of road signs using a neural network", *Innovative Technologies and Scientific Solutions for Industries,* No. 2 (8), P. 130–141. DOI: https://doi.org/10.30837/2522-9818.2019.8.130

3. 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

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. Klin, B., Podpora, M., Beniak, R., Gardecki, A., Rut, J. (2023), "Smart Beamforming in Verbal Human-machine Interaction for Humanoid Robots", *IEEE Robotics and Automation Letters,* P. 4689–4696. DOI: 10.1109/LRA.2023.3288381

7. Jin, R., Ablimit, M., Hamdulla, A. (2023), "Speaker Verification based on Single Channel Speech Separation", *IEEE Access,* available at: https://ieeexplore.ieee.org/iel7/6287639/6514899/10156847.pdf

8. Froiz-Míguez, I., Fraga-Lamas, P., Fernández-Caramés, T. M. (2023), "Design, Implementation and Practical Evaluation of a Voice Recognition Based IoT Home Automation System for Low-Resource Languages and Resource-Constrained Edge IoT Devices: a System for Galician and Mobile Opportunistic Scenarios", *IEEE Access*, available at: https://www.researchgate.net/profile/Tiago-Fernandez-Carames

9. Tesema, F. B., Gu, J., Song, W., Wu, H., Zhu, S., Lin, Z. (2023), "Efficient Audiovisual Fusion for Active Speaker Detection", *IEEE Access*, Vol. 11, P. 45140–45153. DOI: 10.1109/ACCESS.2023.3267668

10. Hu, Z., LingHu, K., Liao, C., Yu, H. (2023), "Speech Emotion Recognition Based on Attention MCNN Combined With Gender Information", *IEEE Access,* Vol. 11, P. 50285–50294. DOI: 10.1109/ACCESS.2023.3278106

11. Barkovska, O., Kholiev, V., Pyvovarova, D., Ivaschenko, G., Rosinskiy, D. (2021), "International system of knowledge exchange for young scientists", *Advanced Information Systems,* No. 5 (1), P. 69–74. DOI: https://doi.org/10.20998/2522-9052.2021.1.09

12. 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

13. Nagrani, A., Chung, J. S., Zisserman, A. (2017), "VoxCeleb: A Large-Scale Speaker Identification Dataset", *Proc. Interspeech 2017,* P. 2616–2620. DOI: https://doi.org/10.21437/Interspeech.2017-950

14. Chung, J. S., Nagrani, A., Zisserman, A. (2018), "VoxCeleb2: Deep Speaker Recognition", *Proc. Interspeech 2018,* P. 1086–1090. DOI: https://doi.org/10.21437/Interspeech.2018-1929

15. Koluguri, N. R., Park, T., Ginsburg, B. (2021), "TitaNet: Neural Model for Speaker Representation with 1D Depth-Wise Separable Convolutions and Global Context", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* P. 8102–8106. DOI: https://doi.org/10.48550/arXiv.2110.04410

16. Koluguri, N. R., Li, J., Lavrukhin, V., Ginsburg, B. (2020), "SpeakerNet: 1D Depth-wise Separable Convolutional Network for Text-Independent Speaker Recognition and Verification", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* DOI: https://doi.org/10.48550/arXiv.2010.12653

17. Dawalatabad, N., Ravanelli, M., Grondin, F., Thienpondt, J., Desplanques, B., Na, H. (2021), "ECAPA-TDNN Embeddings for Speaker Diarization", *Proc. Interspeech, 2021,* P. 3560–3564. DOI: https://doi.org/10.21437/interspeech.2021-941

18. Li, J., Lavrukhin, V., Ginsburg, B., Leary, R., Kuchaiev, O., Cohen, J., Nguyen, H., Gadde, R. (2019), "Jasper: An End-to-End Convolutional Neural Acoustic Model", *Electrical Engineering and Systems Science*, P. 71–75. DOI: https://doi.org/10.21437/Interspeech.2019-1819

Received 16.06.2023

Відомості про авторів / About the Authors

Холєв Владислав Олександрович – Харківський національний університет радіоелектроніки, асистент кафедри електронних обчислювальних машин, Харків, Україна; e-mail: vladyslav.kholiev@nure.ua; ORCID ID: https://orcid.org/0000-0002-9148-1561

Барковська Олеся Юріївна – кандидат технічних наук, доцент, Харківський національний університет радіоелектроніки, доцент кафедри електронних обчислювальних машин, Харків, Україна; e-mail: olesia.barkovska@nure.ua; ORCID ID: https://orcid.org/0000-0001-7496-4353

Kholiev Vladyslav – Kharkiv National University of Radio Electronics, postgraduate at the Department of Electronic Computers, Kharkiv, Ukraine.

Barkovska Olesia – PhD (Engineering Sciences), Associate Professor, Kharkiv National University of Radio Electronics, Associate Professor at the Department of Electronic Computers, Kharkiv, Ukraine.

ПОРІВНЯЛЬНИЙ АНАЛІЗ НЕЙРОМЕРЕЖНИХ МОДЕЛЕЙ ДЛЯ РОЗВ’ЯЗАННЯ ЗАВДАНЬ РОЗПІЗНАВАННЯ СПІКЕРА

Предметом дослідження є нейромережні моделі, розроблені або адаптовані для розв’язання проблеми аналізу голосу в контексті завдань ідентифікації та верифікації спікера. **Метою роботи** є проведення порівняльного аналізу відповідних нейромережних моделей для визначення однієї (або кількох), що якнайкраще відповідає таким обраним критеріям: тип моделі, мова програмування реалізації моделі, потенціал розпаралелювання, чи є модель бінарна, чи мультикласова, точність та обчислювальна складність. Деякі з цих критеріїв обрані, оскільки є універсально важливими, незалежними від того чи іншого завдання, наприклад точність і обчислювальна складність. Інші критерії обрані у зв’язку з архітектурою та недоліками системи наукової комунікації, що виконує завдання ідентифікації та перевірки спікера. **Актуальність** роботи полягає в поширенні аудіо як комунікативного засобу, зокрема йдеться про практичне застосування його інтелектуального аналізу в різних сферах людської діяльності (бізнес, право, військова справа). Крім того, постає питання про необхідність створення ефективного середовища внутрішньої наукової комунікації на основі аудіо серед молодих учених, що дасть їм змогу прискорити свої дослідження й набути навичок наукового спілкування. Для досягнення мети в роботі розв’язані такі **завдання**: сформульовано критерії для оцінюваних моделей з огляду на конкретні потреби й завдання; за певними критеріями досліджено моделі, розроблені для завдань ідентифікації та верифікації спікера. **Результати:** розглянуто моделі *SincNet, VGGVox, Jasper, TitaNet, SpeakerNet, ECAPA_TDNN*; результати дослідження нейромережних моделей зведено в загальну таблицю; визначено оптимальні моделі відповідно до сформульованих критеріїв. **Висновки**: для майбутніх досліджень і практичного розв’язання проблеми автентифікації спікера доцільно використовувати згорткову нейронну мережу, реалізовану мовою програмування *Python*, оскільки вона пропонує широкий вибір інструментів розроблення та бібліотек.

Ключові слова: порівняльний аналіз; нейронна мережа; інтелектуальні моделі; модель; машинне навчання; идентифікація спікера; розпізнавання спікера.

Бібліографічні описи / Bibliographic descriptions

Холєв В. О., Барковська О. Ю. Порівняльний аналіз нейромережних моделей для розв’язання завдань розпізнавання спікера. *Сучасний стан наукових досліджень та технологій в промисловості*. 2023. № 2 (24). С. 172–178. DOI: https://doi.org/10.30837/ITSSI.2023.24.172

Kholiev, V., Barkovska, O. (2023), "Comparative analysis of neural network models for the problem of speaker recognition", *Innovative Technologies and Scientific Solutions for Industries,* No. 2 (24), P. 172–178. DOI: https://doi.org/10.30837/ITSSI.2023.24.172