ISSN 2500-316X (Online)

RUSSIAN TECHNOLOGICAL JOURNAL

РОССИЙСКИЙ
ТЕХНОЛОГИЧЕСКИЙ
ЖУРНАЛ

Information systems.
Computer sciences.
Issues of information security

Multiple robots (robotic centers) and systems.
Remote sensing and nondestructive testing

Modern radio engineering and telecommunication systems

Micro- and nanoelectronics.
Condensed matter physics

Analytical instrument engineering and technology

Mathematical modeling

Economics of knowledge-intensive and high-tech enterprises and industries.
Management in organizational systems

Product quality management. Standardization

Philosophical foundations of technology and society

12+
www.rtj-mirea.ru

13(3) 2025

ISSN 2500-316X (Online)

RUSSIAN TECHNOLOGICAL JOURNAL

РОССИЙСКИЙ ТЕХНОЛОГИЧЕСКИЙ ЖУРНАЛ

- Information systems. Computer sciences. Issues of information security
- Multiple robots (robotic centers) and systems. Remote sensing and nondestructive testing
- Modern radio engineering and telecommunication systems
- Micro- and nanoelectronics. Condensed matter physics
- Analytical instrument engineering and technology
- Mathematical modeling
- Economics of knowledge-intensive and high-tech enterprises and industries. Management in organizational systems
- Product quality management. Standardization
- Philosophical foundations of technology and society

- Информационные системы. Информатика. Проблемы информационной безопасности
- Роботизированные комплексы и системы. Технологии дистанционного зондирования и неразрушающего контроля
- Современные радиотехнические и телекоммуникационные системы
- Микро- и наноэлектроника. Физика конденсированного состояния
- Аналитическое приборостроение и технологии
- Математическое моделирование
- Экономика наукоемких и высокотехнологичных предприятий и производств. Управление в организационных системах
- Управление качеством продукции. Стандартизация
- Мировоззренческие основы технологии и общества

Russian Technological Journal
2025, Vol. 13, No. 3

Russian Technological Journal
2025, том 13, № 3

https://www.rtj-mirea.ru

https://doi.org/10.32362/2500-316X

ISSN 2500-316X (Online)

Russian Technological Journal 2025, Vol. 13, No. 3

Publication date May 30, 2025.

The peer-reviewed scientific and technical journal highlights the issues of complex development of radio engineering, telecommunication and information systems, electronics and informatics, as well as the results of fundamental and applied interdisciplinary researches, technological and economical developments aimed at the development and improvement of the modern technological base.

Periodicity: bimonthly.
The journal was founded in December 2013. The titles were «Herald of MSTU MIREA» until 2016 (ISSN 2313-5026) and «Rossiiskii tekhnologicheskii zhurnal» from January 2016 until July 2021 (ISSN 2500-316X).

Founder and Publisher:
Federal State Budget
Educational Institution of Higher Education
«MIREA – Russian Technological University»
78, Vernadskogo pr., Moscow, 119454 Russia.

The journal is included into the List of peer-reviewed science press of the State Commission for Academic Degrees and Titles of Russian Federation. The Journal is included in Russian Science Citation Index (RSCI), Russian State Library (RSL), Science Index, eLibrary, Directory of Open Access Journals (DOAJ), Directory of Open Access Scholarly Resources (ROAD), Google Scholar, Ulrich's International Periodicals Directory.

Editor-in-Chief:
Alexander S. Sigov, Academician at the Russian Academy of Sciences, Dr. Sci. (Phys.–Math.), Professor, President of MIREA – Russian Technological University (RTU MIREA), Moscow, Russia.
Scopus Author ID 35557510600, ResearcherID L-4103-2017, sigov@mirea.ru.

Editorial staff:

Managing Editor	Cand. Sci. (Eng.) Galina D. Seredina
Scientific Editor	Dr. Sci. (Eng.), Prof. Gennady V. Kulikov
Executive Editor	Anna S. Alekseenko
Technical Editor	Darya V. Trofimova

86, Vernadskogo pr., Moscow, 119571 Russia.
Phone: +7 (499) 600-80-80 (#31288).
E-mail: seredina@mirea.ru.

The registration number ПИ № ФС 77 - 81733 was issued in August 19, 2021 by the Federal Service for Supervision of Communications, Information Technology, and Mass Media of Russia.

The subscription index of *Pressa Rossii*: **79641**.

Russian Technological Journal 2025, том 13, № 3

Дата опубликования 30 мая 2025 г.

Научно-технический рецензируемый журнал освещает вопросы комплексного развития радиотехнических, телекоммуникационных и информационных систем, электроники и информатики, а также результаты фундаментальных и прикладных междисциплинарных исследований, технологических и организационно-экономических разработок, направленных на развитие и совершенствование современной технологической базы.

Периодичность: один раз в два месяца.
Журнал основан в декабре 2013 года. До 2016 г. издавался под названием «Вестник МГТУ МИРЭА» (ISSN 2313-5026), а с января 2016 г. по июль 2021 г. под названием «Российский технологический журнал» (ISSN 2500-316X).

Учредитель и издатель:
федеральное государственное бюджетное
образовательное учреждение высшего образования
«МИРЭА – Российский технологический университет»
119454, РФ, г. Москва, пр-т Вернадского, д. 78.

Журнал входит в Перечень ведущих рецензируемых научных журналов ВАК РФ, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук и доктора наук, входит в RSCI, РГБ, РИНЦ, eLibrary, Directory of Open Access Journals (DOAJ), Directory of Open Access Scholarly Resources (ROAD), Google Scholar, Ulrich's International Periodicals Directory.

Главный редактор:
Сигов Александр Сергеевич, академик РАН, доктор физ.-мат. наук, профессор, президент ФГБОУ ВО МИРЭА – Российский технологический университет (РТУ МИРЭА), Москва, Россия.
Scopus Author ID 35557510600, ResearcherID L-4103-2017, sigov@mirea.ru.

Редакция:

Зав. редакцией	к.т.н. Г.Д. Середина
Научный редактор	д.т.н., проф. Г.В. Куликов
Выпускающий редактор	А.С. Алексеенко
Технический редактор	Д.В. Трофимова

119571, г. Москва, пр-т Вернадского, 86, оф. Р-108.
Тел.: +7 (499) 600-80-80 (#31288).
E-mail: seredina@mirea.ru.

Регистрационный номер и дата принятия решения о регистрации СМИ ПИ № ФС 77 - 81733 от 19.08.2021 г. СМИ зарегистрировано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций (Роскомнадзор).

Индекс по объединенному каталогу «Пресса России» **79641**.

https://www.rtj-mirea.ru

© MIREA – Russian Technological University, 2025

Editorial Board

Stanislav A. Kudzh Dr. Sci. (Eng.), Professor, Rector of RTU MIREA, Moscow, Russia. Scopus Author ID 56521711400, ResearcherID AAG-1319-2019, https://orcid.org/0000-0003-1407-2788, rector@mirea.ru

Juras Banys Habilitated Doctor of Sciences, Professor, Vice-Rector of Vilnius University, Vilnius, Lithuania. Scopus Author ID 7003687871, juras.banys@ff.vu.lt

Vladimir B. Betelin Academician at the Russian Academy of Sciences (RAS), Dr. Sci. (Phys.-Math.), Professor, Supervisor of Scientific Research Institute for System Analysis, RAS, Moscow, Russia. Scopus Author ID 6504159562, ResearcherID J-7375-2017, betelin@niisi.msk.ru

Alexei A. Bokov Dr. Sci. (Phys.-Math.), Senior Research Fellow, Department of Chemistry and 4D LABS, Simon Fraser University, Vancouver, British Columbia, Canada. Scopus Author ID 35564490800, ResearcherID C-6924-2008, http://orcid.org/0000-0003-1126-3378, abokov@sfu.ca

Sergey B. Vakhrushev Dr. Sci. (Phys.–Math.), Professor, Head of the Laboratory of Neutron Research, A.F. Ioffe Physico-Technical Institute of the RAS, Department of Physical Electronics of St. Petersburg Polytechnic University, St. Petersburg, Russia. Scopus Author ID 7004228594, ResearcherID A-9855-2011, http://orcid.org/0000-0003-4867-1404, s.vakhrushev@mail.ioffe.ru

Yury V. Gulyaev Academician at the RAS, Dr. Sci. (Phys.–Math.), Professor, Academic Supervisor of V.A. Kotelnikov Institute of Radio Engineering and Electronics of the RAS, Moscow, Russia. Scopus Author ID 35562581800, gulyaev@cplire.ru

Dmitry O. Zhukov Dr. Sci. (Eng.), Professor of the Department of Telecommunications, Institute of Radio Electronics and Informatics, RTU MIREA, Moscow, Russia. Scopus Author ID 57189660218, zhukov_do@mirea.ru

Alexey V. Kimel PhD (Phys.–Math.), Professor, Radboud University, Nijmegen, Netherlands, Scopus Author ID 6602091848, ResearcherID D-5112-2012, a.kimel@science.ru.nl

Sergey O. Kramarov Dr. Sci. (Phys.–Math.), Professor, Surgut State University, Surgut, Russia. Scopus Author ID 56638328000, ResearcherID E-9333-2016, https://orcid.org/0000-0003-3743-6513, mavoo@yandex.ru

Dmitry A. Novikov Academician at the RAS, Dr. Sci. (Eng.), Director of V.A. Trapeznikov Institute of Control Sciences, Moscow, Russia. Scopus Author ID 7102213403, ResearcherID Q-9677-2019, https://orcid.org/0000-0002-9314-3304, novikov@ipu.ru

Philippe Pernod Dr. Sci. (Electronics), Professor, Dean of Research of Centrale Lille, Villeneuve-d'Ascq, France. Scopus Author ID 7003429648, philippe.pernod@ec-lille.fr

Mikhail P. Romanov Dr. Sci. (Eng.), Professor, Academic Supervisor of the Institute of Artificial Intelligence, RTU MIREA, Moscow, Russia. Scopus Author ID 14046079000, https://orcid.org/0000-0003-3353-9945, m_romanov@mirea.ru

Viktor P. Savinykh Academician at the RAS, Dr. Sci. (Eng.), Professor, President of Moscow State University of Geodesy and Cartography, Moscow, Russia. Scopus Author ID 56412838700, vp@miigaik.ru

Andrei N. Sobolevski Professor, Dr. Sci. (Phys.-Math.), Director of Institute for Information Transmission Problems (Kharkevich Institute), Moscow, Russia. Scopus Author ID 7004013625, ResearcherID D-9361-2012, http://orcid.org/0000-0002-3082-5113, sobolevski@iitp.ru

Li Da Xu Academician at the European Academy of Sciences, Russian Academy of Engineering (formerly, USSR Academy of Engineering), and Armenian Academy of Engineering, Dr. Sci. (Systems Science), Professor and Eminent Scholar in Information Technology and Decision Sciences, Old Dominion University, Norfolk, VA, the United States of America. Scopus Author ID 13408889400, https://orcid.org/0000-0002-5954-5115, lxu@odu.edu

Yury S. Kharin Academician at the National Academy of Sciences of Belarus, Dr. Sci. (Phys.–Math.), Professor, Director of the Institute of Applied Problems of Mathematics and Informatics of the Belarusian State University, Minsk, Belarus. Scopus Author ID 6603832008, http://orcid.org/0000-0003-4226-2546, kharin@bsu.by

Yuri A. Chaplygin Academician at the RAS, Dr. Sci. (Eng.), Professor, Member of the Departments of Nanotechnology and Information Technology of the RAS, President of the National Research University of Electronic Technology (MIET), Moscow, Russia. Scopus Author ID 6603797878, ResearcherID B-3188-2016, president@miet.ru

Vasily V. Shpak Cand. Sci. (Econ.), Deputy Minister of Industry and Trade of the Russian Federation, Ministry of Industry and Trade of the Russian Federation, Moscow, Russia; Associate Professor, National Research University of Electronic Technology (MIET), Moscow, Russia, mishinevaiv@minprom.gov.ru

Редакционная коллегия

Кудж **Станислав Алексеевич**	д.т.н., профессор, ректор РТУ МИРЭА, Москва, Россия. Scopus Author ID 56521711400, ResearcherID AAG-1319-2019, https://orcid.org/0000-0003-1407-2788, rector@mirea.ru
Банис **Юрас Йонович**	хабилитированный доктор наук, профессор, проректор Вильнюсского университета, Вильнюс, Литва. Scopus Author ID 7003687871, juras.banys@ff.vu.lt
Бетелин **Владимир Борисович**	академик Российской академии наук (РАН), д.ф.-м.н., профессор, научный руководитель Федерального научного центра «Научно-исследовательский институт системных исследований» РАН, Москва, Россия. Scopus Author ID 6504159562, ResearcherID J-7375-2017, betelin@niisi.msk.ru
Боков **Алексей Алексеевич**	д.ф.-м.н., старший научный сотрудник, химический факультет и 4D LABS, Университет Саймона Фрейзера, Ванкувер, Британская Колумбия, Канада. Scopus Author ID 35564490800, ResearcherID C-6924-2008, http://orcid.org/0000-0003-1126-3378, abokov@sfu.ca
Вахрушев **Сергей Борисович**	д.ф.-м.н., профессор, заведующий лабораторией нейтронных исследований Физико-технического института им. А.Ф. Иоффе РАН, профессор кафедры Физической электроники СПбГПУ, Санкт-Петербург, Россия. Scopus Author ID 7004228594, ResearcherID A-9855-2011, http://orcid.org/0000-0003-4867-1404, s.vakhrushev@mail.ioffe.ru
Гуляев **Юрий Васильевич**	академик РАН, д.ф.-м.н., профессор, научный руководитель Института радиотехники и электроники им. В.А. Котельникова РАН, Москва, Россия. Scopus Author ID 35562581800, gulyaev@cplire.ru
Жуков **Дмитрий Олегович**	д.т.н., профессор кафедры телекоммуникаций Института радиоэлектроники и информатики РТУ МИРЭА, Москва, Россия. Scopus Author ID 57189660218, zhukov_do@mirea.ru
Кимель **Алексей Вольдемарович**	к.ф.-м.н., профессор, Университет Радбауд, г. Наймеген, Нидерланды. Scopus Author ID 6602091848, ResearcherID D-5112-2012, a.kimel@science.ru.nl
Крамаров **Сергей Олегович**	д.ф.-м.н., профессор, Сургутский государственный университет, Сургут, Россия. Scopus Author ID 56638328000, ResearcherID E-9333-2016, https://orcid.org/0000-0003-3743-6513, mavoo@yandex.ru
Новиков **Дмитрий Александрович**	академик РАН, д.т.н., директор Института проблем управления им. В.А. Трапезникова РАН, Москва, Россия. Scopus Author ID 7102213403, ResearcherID Q-9677-2019, https://orcid.org/0000-0002-9314-3304, novikov@ipu.ru
Перно Филипп	Dr. Sci. (Electronics), профессор, Центральная Школа г. Лилль, Франция. Scopus Author ID 7003429648, philippe.pernod@ec-lille.fr
Романов **Михаил Петрович**	д.т.н., профессор, научный руководитель Института искусственного интеллекта РТУ МИРЭА, Москва, Россия. Scopus Author ID 14046079000, https://orcid.org/0000-0003-3353-9945, m_romanov@mirea.ru
Савиных **Виктор Петрович**	академик РАН, Дважды Герой Советского Союза, д.т.н., профессор, президент Московского государственного университета геодезии и картографии, Москва, Россия. Scopus Author ID 56412838700, vp@miigaik.ru
Соболевский **Андрей Николаевич**	д.ф.-м.н., директор Института проблем передачи информации им. А.А. Харкевича, Москва, Россия. Scopus Author ID 7004013625, ResearcherID D-9361-2012, http://orcid.org/0000-0002-3082-5113, sobolevski@iitp.ru
Сюй **Ли Да**	академик Европейской академии наук, Российской инженерной академии и Инженерной академии Армении, Dr. Sci. (Systems Science), профессор, Университет Олд Доминион, Норфолк, Соединенные Штаты Америки. Scopus Author ID 13408889400, https://orcid.org/0000-0002-5954-5115, lxu@odu.edu
Харин **Юрий Семенович**	академик Национальной академии наук Беларуси, д.ф.-м.н., профессор, директор НИИ прикладных проблем математики и информатики Белорусского государственного университета, Минск, Беларусь. Scopus Author ID 6603832008, http://orcid.org/0000-0003-4226-2546, kharin@bsu.by
Чаплыгин **Юрий Александрович**	академик РАН, д.т.н., профессор, член Отделения нанотехнологий и информационных технологий РАН, президент Института микроприборов и систем управления им. Л.Н. Преснухина НИУ «МИЭТ», Москва, Россия. Scopus Author ID 6603797878, ResearcherID B-3188-2016, president@miet.ru
Шпак **Василий Викторович**	к.э.н., зам. министра промышленности и торговли Российской Федерации, Министерство промышленности и торговли РФ, Москва, Россия; доцент, Институт микроприборов и систем управления им. Л.Н. Преснухина НИУ «МИЭТ», Москва, Россия, mishinevaiv@minprom.gov.ru

Contents

Information systems. Computer sciences. Issues of information security

Multiple robots (robotic centers) and systems. Remote sensing and non-destructive testing

Modern radio engineering and telecommunication systems

Micro- and nanoelectronics. Condensed matter physics

Содержание

Информационные системы. Информатика. Проблемы информационной безопасности

Роботизированные комплексы и системы. Технологии дистанционного зондирования неразрушающего контроля

Современные радиотехнические и телекоммуникационные системы

Микро- и наноэлектроника. Физика конденсированного состояния

Information systems. Computer sciences. Issues of information security

Информационные системы. Информатика. Проблемы информационной безопасности

UDC 004.522
https://doi.org/10.32362/2500-316X-2025-13-3-7-20
EDN PVYBDD

RESEARCH ARTICLE

Accent conversion method with real-time voice cloning based on a non-autoregressive neural network model

Vladimir A. Nechaev @,
Sergey V. Kosyakov

Ivanovo State Power Engineering University, Ivanovo, 153003 Russia
@ *Corresponding author, e-mail: nechaev@gapps.ispu.ru*

• **Submitted:** 23.07.2024 • **Revised:** 03.02.2025 • **Accepted:** 26.03.2025

Abstract

Objectives. The development of contemporary models for the conversion of accents in foreign languages utilizes deep neural network architectures, as well as ensembles of neural networks for speech recognition and generation. However, restricted access to implementations of such models limits their application, study, and further development. Moreover, the use of these models is limited by their architectural features, which prevents flexible changes from being carried out in the timbre of the generated speech and requires the accumulation of context, leading to increased delays in generation, making these systems unsuitable for use in real-time multi-user communication scenarios. Therefore, the relevant task and aim of this work is the development of a method that generates native-sounding speech based on input accented speech material with minimal delays and the capability to preserve, clone, and modify the timbre of the speaker's voice.

Methods. Methods for modifying, training, and combining deep neural networks into a single end-to-end architecture for direct speech-to-speech conversion are applied. For training, original and modified open-source datasets were used.

Results. The work resulted in the development of a real-time accent conversion method with voice cloning based on a non-autoregressive neural network. The model comprises modules for accent and gender detection, speaker identification, speech conversion, spectrogram generation, and decoding the resulting spectrogram into an audio signal. As well as demonstrating high accent conversion quality while maintaining the original timbre, the short generation times of the applied method make it acceptable for use in real-time scenarios.

Conclusions. Testing of the developed method confirmed the effectiveness of the proposed non-autoregressive neural network architecture. The developed model demonstrated the ability to work in real-time information systems in English.

Keywords: accent conversion, speech synthesis, text-to-speech, voice conversion, machine learning, neural network

For citation: Nechaev V.A., Kosyakov S.V. Accent conversion method with real-time voice cloning based on a non-autoregressive neural network model. *Russian Technological Journal.* 2025;13(3):7–20. https://doi.org/10.32362/2500-316X-2025-13-3-7-20, https://www.elibrary.ru/PVYBDD

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

© V.A. Nechaev, S.V. Kosyakov, 2025

НАУЧНАЯ СТАТЬЯ

Метод конвертации акцента с клонированием голоса в реальном времени на основе неавторегрессионной нейросетевой модели

В.А. Нечаев [@]**,**
С.В. Косяков

Ивановский государственный энергетический университет имени В.И. Ленина, Иваново, 153003 Россия
[@] *Автор для переписки, e-mail: nechaev@gapps.ispu.ru*

• **Поступила:** 23.07.2024 • **Доработана:** 03.02.2025 • **Принята к опубликованию:** 26.03.2025

Резюме

Цели. В настоящее время при разработке моделей для преобразования речи с акцентом в речь без акцента используются архитектуры глубоких нейросетей, а также ансамбли предобученных нейросетей для распознавания и генерации речи. При этом доступ к реализациям таких моделей является ограниченным, что затрудняет их применение, изучение и дальнейшее развитие. Также использование данных моделей ограничено особенностями архитектуры, которая не позволяет гибко менять тембр генерируемой речи и требует накопления контекста, что ведет к увеличению задержки при генерации и делает данные системы непригодными для использования в сценариях коммуникации двух и более людей в реальном времени. В связи с этим актуальной задачей и целью настоящей работы является разработка метода, позволяющего на основе входной речи с акцентом генерировать речь без акцента с минимальными задержками с возможностью сохранения, клонирования и модификации тембра говорящего, что позволит преодолеть ограничения текущих моделей.

Методы. Применены методы модификации, обучения и объединения глубоких нейросетей в единую сквозную архитектуру для прямого преобразования речи в речь. Для обучения использованы оригинальные и модифицированные наборы данных из открытых источников.

Результаты. Разработан метод конвертации акцента с клонированием голоса в реальном времени на основе неавторегрессионной нейросетевой модели, которая состоит из модулей определения акцента и пола, идентификации говорящего, преобразования речи в фонетическое представление, генерации спектрограммы и декодирования полученной спектрограммы в аудиосигнал. Метод демонстрирует высокое качество конвертации акцента с сохранением оригинального тембра, а также низкие задержки при генерации, приемлемые для использования в сценариях реального времени.

Выводы. Апробация разработанного метода подтвердила эффективность предложенной неавторегрессионной нейросетевой архитектуры. Разработанная прикладная нейросетевая модель продемонстрировала возможность работы в информационных системах на английском языке в режиме реального времени.

Ключевые слова: конвертация акцента, генерация речи, распознавание речи, конвертация голоса, машинное обучение, нейронная сеть

Для цитирования: Нечаев В.А., Косяков С.В. Метод конвертации акцента с клонированием голоса в реальном времени на основе неавторегрессионной нейросетевой модели. *Russian Technological Journal*. 2025;13(3):7–20. https://doi.org/10.32362/2500-316X-2025-13-3-7-20, https://www.elibrary.ru/PVYBDD

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

One of the most important channels of interaction between businesses and their customers is communication through voice communication. This is evidenced by the development and widespread use of call centers, which can now be established to operate on a cross-national and cross-regional basis. In such cases, English is very often used to overcome the interlingual barrier despite the operators and clients of call centers not all being native speakers of this language. As a result, situations may arise when a customer abandons a communication due to the difficulty of mutual understanding with the operator, which leads to economic losses. With the development of artificial intelligence systems, accent conversion software systems have been used to solve this problem, which enable to reduce the speaker's accent to a certain extent [1]. Such systems can also be used in the process of teaching foreign languages [2, 3], re-recording and enhancing the quality of previously recorded speech [4], and improving the quality of existing speech recognition systems [5]. Despite the considerable developments that have recently taken place in this area of research, the problem of improving and enhancing the quality of such systems remains relevant.

Accents in speech, representing an integral feature of pronunciation, can be divided into native accents, which depend on many regional and cultural factors, and foreign accents [6]. At the same time, foreign accent differs from native accent at the segmental (phonemes) and suprasegmental (intonation, accents, rhythmics) levels [7]. A foreign accent manifests itself when a native speaker of one language (L1 speech) speaks in another non-native or second language (L2 speech) [8]. L2 speech can be less intelligible to native speakers than L1 speech based on similar content [9], resulting in reduced comprehension of and trust in what is said, negative attitudes towards the speaker, and other forms of discrimination by native speakers [10–12].

Early methods for accent conversion in the generation phase are based on reference L1 examples corresponding to L2 speech [13–16]. For each L2 phrase, a corresponding L1 phrase is required. The practical application of such models is limited due to insufficient data to cover all possible speech variations. Such approaches also require significant resources for data collection and processing, which increases the development time and cost of such systems. In addition, the strict adherence to pairwise examples may reduce the versatility and scalability of the technology by limiting its ability to adapt to new accents or speech styles that were not included in the original dataset.

In subsequent developments, this limitation was overcome, meaning that reference examples are not required at the inference stage [17–22]. However, parallel datasets containing similar L2 and L1 phrases are still used for model training [17, 20], necessitating difficult and expensive operations for obtaining a sufficient amount of such data. Moreover, the autoregressive recurrent neural networks used by the described methods complicates the process of training them. Another method [19] uses pretrained neural networks to convert text to speech. To preserve individual voice characteristics, a separate model would need to be trained for each target speaker, making multi-user use difficult. Methods [21, 22] based on predicting the duration of each generated phoneme, which transforms the original L2 speech duration and speaker identity, require the accumulation of context, increasing generation time and complicating real-time use. Although the method described in [18] is not subject to the disadvantages listed above, the implementation of the model is limited to a finite set of accents, whose identifiers have to be determined during the model training phase. This complicates the process of training the data and applying the model with accents that have not been previously represented in it.

The present work set out to develop an accent conversion method that overcomes the above problems and shortcomings, capable of converting any speech from L2 to L1 without using reference examples or parallel data in the training and generation stages, which greatly simplifies, cheapens and speeds up the process of adapting the system to new accents.

1. RESEARCH OBJECTIVES

Speakers perceive literacy and fluency, on the one hand, and accent, on the other hand, as separate entities; improvement of one of them leads to an enhanced overall perception of L2 speech by native speakers [9, 23]. At the same time, absolutely accurate reproduction of L1 speech by a non-native speaker is difficult to achieve in practice due to differences in phonetic interference and speech perception by native speakers [24]. When solving the task of accent conversion, it is important to preserve the speaker's individual vocal features (timbre, pitch, loudness), i.e., it is necessary to perform voice cloning [25] while modifying segmental and suprasegmental characteristics associated with the foreign accent and pronunciation [13, 17, 18]. This is especially important in situations where it is necessary to preserve the emotional coloring, expressiveness, and individual features of speech, including voice features related to the speaker's gender.

In order to fulfill these conditions, it is necessary to use several interconnected, end-to-end architecture models for accent and gender detection, speaker embedding (SE), speech-to-phonetic (STP) representation, and spectrogram generation, as well as the decoding of the resulting spectrogram into an audio signal.

In addition, in real-time scenarios, such as human voice communication using high-speed communication channels, it is necessary to ensure minimal delays in the generation and transmission of processed speech [26–28]. Accent translation should be performed in real-time without the use of recurrent networks [29] to avoid the error accumulation effect associated with sequential output generation.

The training phase is based on publicly available open data for training the speech recognition and generation systems. In addition, the method set out to be independent of the availability of benchmark examples and parallel data at the training and inference stages.

2. RESEARCH METHODS

2.1. Architecture of the developed system

The developed accent conversion method includes several interrelated models integrated into a single end-to-end architecture for accent and gender detection, SE, STP conversion, spectrogram generation, and decoding of the obtained spectrogram into an audio signal. Figure 1 shows the general scheme of interaction of the above models at the output stage (generation of output L1 audio).

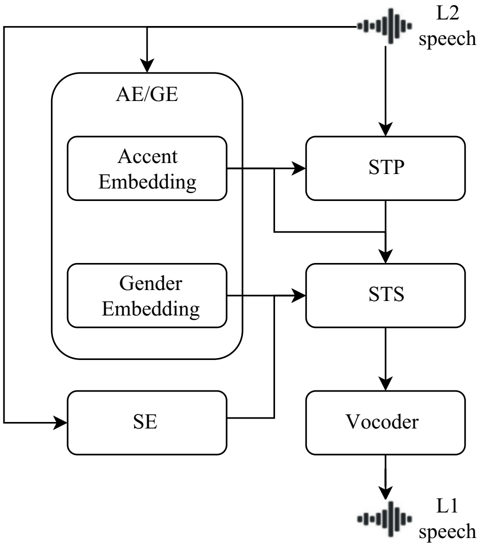


Fig. 1. General derivation scheme of the accent conversion method with voice cloning

The L2 speech audio signal is fed to the input of the STP model, to the input of the Accent and Gender Embedding (AE/GE) model, and to the input of the SE model. The accent embedding affects the generation of the phonetic representation, which in vectorized form is fed to the input of the speech-to-speech (STS) and mel spectrogram generation model. The AE/GE vector representations (embeddings) are also fed to the input of the STP model, as well as the output of the SE model, which is a vector representation of individual voice characteristics (timbre). The resulting spectrogram is converted into an L1 speech audio signal using a decoding vocoder model.

The overall pipeline of L1 speech generation from the original L2 speech can be simplistically represented as a formula:

$$a_{\mathrm{L1}} = F_{\mathrm{V}}(F_{\mathrm{STS}}(F_{\mathrm{STP}}(a_{\mathrm{L2}}, F_{\mathrm{AE}}(a_{\mathrm{L2}})), F_{\mathrm{AE}}(a_{\mathrm{L2}}), F_{\mathrm{GE}}(a_{\mathrm{L2}}), F_{\mathrm{SE}}(a_{\mathrm{L2}}))), \quad (1)$$

where a_{L1} is the generated L1 speech audio signal; a_{L2} is the input L2 speech audio signal; F_{V} is the vocoder model; F_{STS} is the STS model; F_{STP} is the STP model; F_{AE} is AE in the AE/GE model; F_{GE} is GE in the AE/GE model; F_{SE} is the SE model, vector representation of individual voice characteristics.

In order to obtain a single end-to-end model of accent conversion, it is necessary to perform the training process of each model sequentially. Thus, the AE/GE and SE models are independent of other models and their training can be performed in any order. The output of the trained AE/GE model will be required at the stage of obtaining the STP model. All previous models (AE/GE, SE, STP) are required to derive the STS model. The training of the vocoder model is based on the output of the STS model.

2.2. AE/GE model

In order to obtain fixed length vectors representing the accent and gender features of the speaker, the model is first trained for the classification task. In such a configuration, class labels used in the training process are returned by the model at the output of the last layer, while vector representations used as voice features are taken from a special intermediate layer.

This and other models use a preprocessor based on fast Fourier transform, which converts the incoming audio signal (time domain) into a mel spectrogram (frequency domain), showing the frequency content of the audio signal on a perceptual mel scale, which approximates the nonlinear frequency response of the human ear. Here, the sampling frequency (sampling rate) is 22050 Hz and the window width is 1024 sound fragments (samples), while the window step is 256 samples and the number of generated mel bands is 80.

Figure 2 shows the training scheme of the accent and gender detection model. It contains blocks of convolutional network of Jasper architecture of 3 × 3 configuration [30]. The accent decoder and gender decoder, which have a common architecture, consist of an attention pooling layer [31], a normalization layer, a convolutional layer to obtain AE and GE of dimension 192, and a linear layer to obtain (predict) the accent class (AC) and gender class (GC).

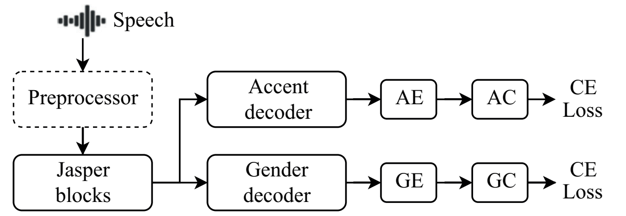


Fig. 2. Training diagram of the AE/GE model. CE are cross entropies

After feeding the audio signal to the preprocessor, the mel spectrogram is fed to the Jasper blocks, as well as, in parallel, to the accent decoder and the gender decoder, along with the corresponding fully connected layers in the output to obtain the accent and gender prediction vectors. During the model training process, the sum of CEs is minimized:

$$L_{\mathrm{AE,GE}}=(x_{\mathrm{a}},y_{\mathrm{a}},x_{\mathrm{g}},y_{\mathrm{g}})=$$

$$=-\sum_{i=1}^{A} y_{\mathrm{a}_i}\ln\left(\frac{\exp x_{\mathrm{a}_i}}{\sum\limits_{k=1}^{A}\exp x_{\mathrm{a}_k}}\right)-\sum_{j=1}^{G} y_{\mathrm{g}_j}\ln\left(\frac{\exp x_{\mathrm{g}_j}}{\sum\limits_{l=1}^{G}\exp x_{\mathrm{g}_l}}\right), \quad (2)$$

where $L_{\mathrm{AE,\ GE}}$ is the overall loss function of the AE/GE model; A is the number of ACs (40); G is the number of GCs (2); x_{a} are accent predictions; x_{g} are gender predictions; y_{a} are ground truth accent labels; y_{g} are ground truth gender labels.

2.3. SE model

Figure 3 shows the training scheme of the SE model and tone vector representation. This scheme contains an input convolutional neural network of SincNet architecture [32], layers of X-Vector DNN[1] model [33], and a layer for obtaining vector representations of dimensionality 512.

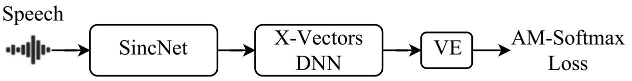


Fig. 3. Diagram of the SE model training. VE is the voice embedding—vector representation of the speaker's individual voice characteristics

Unlike the AE/GE model, the audio signal is not pretransformed into a mel spectrogram, i.e., the digitized audio signal in the time domain with a sampling rate of 16000 Hz is fed to the band-pass filters of the SincNet architecture, then to the convolution layers of the X-Vector DNN and the output fully connected layer, which provides a vector representation of individual voice characteristics (timbre) at the output. In the process of model training, the problem of representation learning is solved while minimizing the Additive Angular Margin (AAM Loss) function [34].

2.4. STP conversion model

The next step is speech recognition taking into account the speaker's accent. For this purpose, it is necessary to obtain a model for converting speech into phonetic or textual representation. The training scheme of the STP model is shown in Fig. 4. The dotted line represents the blocks that are fixed (frozen) during the backpropagation process, i.e., the weights in these blocks are not updated, but their previously obtained states are used.

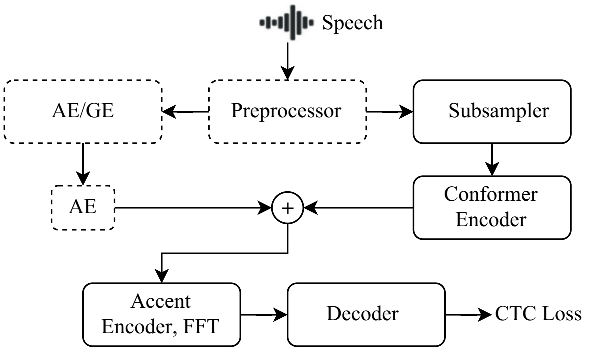


Fig. 4. Diagram of STP model training

The speech audio signal is fed to the input of the preprocessor as previously described and then in parallel to the AE/GE model to obtain the AE and to the convolutional dimensionality reduction block (Subsampler) with a factor of 4. Further conversion is carried out using the Conformer encoder, which comprises a 12 module Conformer architecture [35] having an internal dimensionality of 512 consisting of fully connected [36], convolutional [37], and attention mechanism transformer layers [38]. AE is then normalized, reduced to dimensionality 512, summed with the output of the Conformer encoder, and fed to the input of the accent encoder, which has a single stack feed-forward transformer (FFT) architecture [39]. The output of the accent encoder is further utilized in the STS model as a distribution of phonetic tokens. Finally, the output signal of the accent encoder is fed to the decoder, which has a single-layer convolutional architecture with Softmax activation function, and forms at the output a vector of predictions of textual tokens of dimension equal to the size of the tokenizer dictionary (128) plus one (for a blank token). During model training, the Connectionist Temporal Classification (CTC) Loss function [40] is minimized, which calculates the loss between the continuous (unsegmented) time series and the target sequence:

[1] Deep neural network.

$$L_{\mathrm{STP}}(x,y) = -\ln\left(\sum_{\rho\in A_{x,y}} \prod_{t=1}^{T} x_{\rho_t}\right), \tag{3}$$

where L_{STP} is the loss function of the STP model (CTC Loss); x is the probabilities of text tokens predicted by the model; y is the sequence of text tokens from the target text; ρ is the alignment path x predictions to reduce to y sequence by removing all blank tokens and merging repeated tokens; $A_{x,y}$ is the set of all possible alignment paths; T is the number of predicted tokens in x; x_{ρ_t} is the probability of a particular predicted token at step t given the chosen alignment path ρ.

2.5. STS conversion and spectrogram generation model

The previous models are combined into a single architecture for STS conversion and spectrogram generation. Figure 5 presents a schematic of its training. The STS model includes the previously discussed blocks of preprocessor, accent, gender (AE/GE model) and speaker's timbre (SE model) with the corresponding modules of vector representations (AE, GE, VE), as well as the STP conversion block (STP model). All these blocks are marked with a dotted line due to their training was performed earlier and is not performed at the stage of STP-model training. Moreover, an untrained block based on normalized cross correlation function and median smoothing was added to the architecture to extract the fundamental or lowest frequency of the periodic sound signal (F0), which is perceived by the human ear as pitch [41, 42].

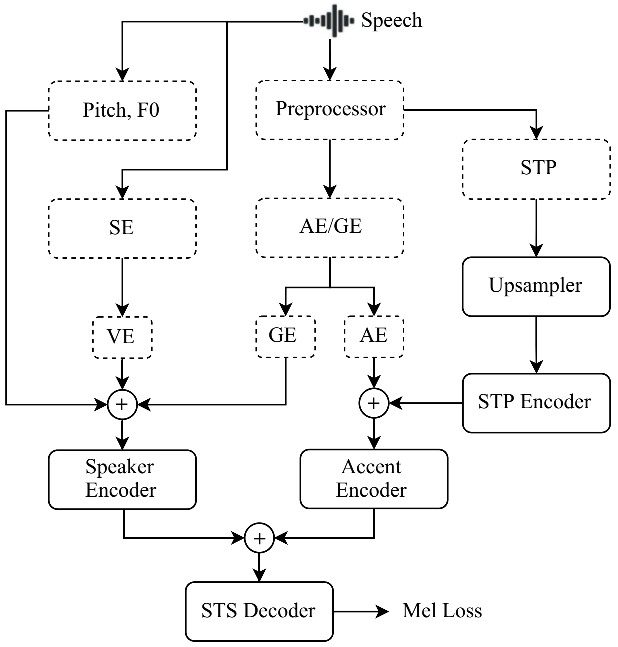


Fig. 5. STS Model training diagram

The speech audio signal is fed to the preprocessor, pitch block and SE model input. The mel spectrogram from the preprocessor is fed to the inputs of the AE/GE and STP models. The phonetic representation from the STP model is fed to an upsampler with a factor of 4 to equalize the original and generated spectrograms, consisting of two convolutional 1D-transposed layers and two rectified linear unit (ReLU) activation functions placed after each convolutional layer. After the upsampler, the phonetic representation is transformed using a STP encoder, which has a six-stack feed-forward transformer (FFT) architecture [39] used in the Fastpitch architecture as an input unit operating in the token domain [43], with inner and outer dimensions of 1536 and 384, respectively. The vector representations of accent, gender, speaker's timbre and pitch profile are normalized and reduced to dimensionality 384. Next, the accent vectors and the output of the STP encoder are summed and fed to the input of the accent encoder (1 stack FFT). Similarly, the pitch, timbre, gender vectors are summed and fed to the input of the speaker encoder (1 stack FFT). Thus, the speaker encoder aggregates speech properties related to individual voice characteristics except accent, which in turn is the responsibility of the accent encoder. The sum of the output vectors of the speaker encoder and the accent encoder is fed to the input of a STS decoder consisting of 6 stacks of Fastpitch architecture FFTs from the output mel area [43]. Finally, the vector is projected to dimension 80 to match the original number of mel areas. During the training process, the loss function is minimized based on the standard deviation:

$$L_{\mathrm{STS}}(x,y) = \frac{1}{\sum_{i=1}^{N} d_i} \sum_{i=1}^{N} d_i (y_i - x_i)^2, \tag{4}$$

where L_{STS} is the loss function of the STS model (Mel Loss); N is the number of elements in the mel spectrogram; x is the mel spectrogram predicted by the model; y is the target mel spectrogram; d is the mask of the spectrogram duration for collection into a batch of fixed size, consisting of values 1 (the element should be considered) and 0 (the element should not be considered), obtained from the duration of the predicted spectrogram.

2.6. Model of sound signal generation from mel spectrogram (vocoder)

Mel spectrogram of L1 speech in the frequency domain obtained using the STS model is converted into a sound signal in the time domain. For this purpose, a model based on generative-adversarial networks (HiFi-GAN) [44] is used. The output audio signal has a sampling rate of 22050 Hz. The model is trained as follows: the audio signal from the training dataset is converted into a mel spectrogram using the STS model, then the resulting spectrogram is passed to a vocoder and converted into an audio signal. Using the received and original audio signals, the loss functions for the generator and discriminator are calculated as described in [44].

2.7. STS conversion and spectrogram generation simplified model (Ablation)

In order to conduct comparative experiments, a simplified version of the accent conversion model was also developed, the schematic of which is shown in Fig. 6.

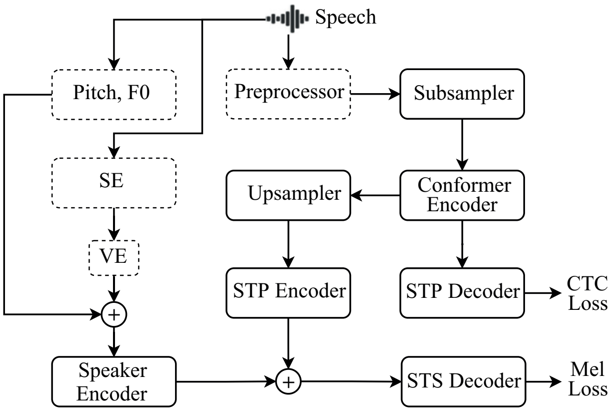


Fig. 6. Schematic of a simplified accent conversion model

This simplified model excludes the AE/GE model, as well as all related encoders in the STP and STS models. Thus, in the resulting simplified model, the output is not conditioned on accent and gender properties. In addition, the training of the STP model was conducted not separately, but simultaneously with the STS model without fixing the weights of the STP with minimization of the sum of the CTC Loss and Mel Loss functions.

3. PRACTICAL APPLICATION OF THE METHOD

3.1. Model training

AE/GE model was trained on the following datasets: CMU-ARCTIC [45], L2 ARCTIC [46], Speech Accent Archive [47], Common Voice 16.1 [48]. All of them represent audio recordings of speech in English, their corresponding textual transcriptions, and contain additional meta-information about accent, gender and, in some cases, native language, place of residence and age of the speaker. Using this information, the audio files were grouped into 40 classes denoting native or foreign English accents, e.g., British, American, Russian, Indian and South Asian, Canadian, German, Australian, African, Japanese, Eastern European, etc. The gender of the speaker is also highlighted. The total duration of the audio files marked in this way amounted to 1087.6 h for the training sample, and 7.6 h for the validation and test samples.

VoxCeleb1 [49] and VoxCeleb2 [50] data collections with a total duration of 2794 h have been used to train the SE model. These sets represent grouped speech audio recordings of 7363 individuals. Audio recordings pertaining to one person are presented during training as positive examples and, conversely, those pertaining to different people as negative examples.

STP model was trained on data from CMU-ARCTIC [45], L2 ARCTIC [46], Common Voice 16.1 [48], LibriSpeech [51], NPTEL2020[2], VCTK [52], GigaSpeech [53]. The mentioned sets consist of audio recordings of English speech with different accents and corresponding text transcriptions. The total duration of the pooled training sample was 6107 h and the validation sample was 48 h. The text transcriptions were normalized, i.e., converted from the canonical written form to the spoken form [54], which is especially important for numbers and abbreviations, and were also brought to a unified form: lower case into ASCII format, punctuation, special characters and additional indentation were removed. A SentencePiece tokenizer [55] with a dictionary size of 128 was trained on the training part of the texts, with which all the texts are processed during the training and evaluation of the model.

STS model and vocoder were trained using the following datasets: CMU-ARCTIC [45], L2 ARCTIC [46], VCTK [52], LibriTTS-R [56], LJ Speech[3]. When we split the data into training and validation samples, their duration was 681 and 17.6 h, respectively. Only audio information without textual markup is used in the training process.

A simplified model (ablation) was trained on data for the STP and STS models.

In order to train, evaluate, and use the described models, code has been developed using the open-source libraries Pytorch [57] and NVIDIA NeMo [58]. The implementation and weights of the vector representation model of the speaker's timbre (SE model) are taken from the Pyannote library [59]. Training was conducted on a server with 8 NVIDIA Tesla V100 graphics processing units (GPUs).

AE/GE model was trained using the stochastic gradient descent optimizer with a learning rate of $1 \cdot 10^{-3}$, weight decay $2 \cdot 10^{-4}$, momentum 0.9, and a Cosine

[2] NPTEL2020 – Indian English Speech Dataset. https://github.com/AI4Bharat/NPTEL2020-Indian-English-Speech-Dataset. Accessed May 01, 2024.

[3] Ito K., Johnson L. *The LJ Speech Dataset*. https://keithito.com/LJ-Speech-Dataset/. Accessed May 01, 2024.

Annealing scheduler for 200 epochs. To train the STP and STS models, the AdamW optimizer was used with a learning rate of $1 \cdot 10^{-3}$, a regularization factor of 0.001, and a similar scheduler as for the AE/GE model for 50 epochs for each model. Fine-tuning of the HiFi-GAN vocoder model was performed by initializing model weights obtained from open sources [44], using the AdamW optimizer and a learning rate of $1 \cdot 10^{-6}$ for 40 epochs. The training of the simplified model (ablation) was carried out with similar parameters used in the STS model.

Table 1 shows the number of trainable parameters of the models optimized during training. In total, the considered accent conversion architecture (full STS), consisting of several interconnected models, has 164 mln parameters.

Table 1. Number of trained parameters

Model	Number of parameters, mln
AE/GE	24.9
SE	4.3
STP	82.1
STS	52.7
Full STS	**164**
Vocoder	84.7
Total	248.7

3.2. Performance assessment

The model performance was assessed on a Linux server running a single NVIDIA Tesla T4 GPU, an 8-core virtual central processing unit (vCPU), and 16 GB of random-access memory (RAM). To accomplish this, the model was first exported to the open source ONNX format and then deployed using NVIDIA Triton open-source software. Using the program interface of the NVIDIA Triton-deployed model and a 5 s test audio file containing English L2 speech, we measured the response generation latency at 200 iterations. As a result, the average latency was 52 ms and throughput was 96 RTFX.

Performance assessment results of the accent conversion model show low generation delays. Together with the features of the architecture, which does not require the accumulation of long context but can handle segments of less than 0.25 s duration, this makes it possible to apply the proposed model in real-time dialog when response delays affect communication [26–28].

3.3. Objective quality assessment

Open-source data as well as pretrained speech recognition models were used to perform objective quality assessment. Using the proposed accent conversion method, an audio file was generated for each example from the test set. Quality metrics were then calculated for the original and corrected audio files. Table 2 presents the results of the objective quality assessment.

As test datasets, we used subsamples totaling 26.9 h that did not participate in the process of training the accent conversion model and its components. All of them include text transcriptions and audio files of English speech with different native and non-native accents from open sources:

- 3.2 h from CMU-ARCTIC [45], L2 ARCTIC [46] (ARCTIC), 10 accents: American, English, Chinese, Indian, Korean, Vietnamese, Spanish, Arabic, Dutch, German;
- 3.1 h from Common Voice [48], 12 accents: American, English, Indian, Australian, African, Chinese, Filipino, Malaysian, German, Russian, French, Eastern European;
- 15.2 h from NPTEL2020, Indian accent;
- 5.4 h of Afrispeech-200 [60], African accent (Yoruba, Swahili, Igbo, Zulu, Tswana, Idoma, Afrikaans).

Speech recognition models obtained from open sources were used: Conformer [35], Citrinet [61], and Whisper [62]. In this case, the Whisper model is taken in two variants: large multilingual (L. Mult.) and medium English (M. En.). Recognition was performed on audio files without processing and on audio files after accent conversion. The recognized and true transcriptions were then reduced to a single form using normalization [54], after which quality metrics were compared and counted: word error rate (WER), character error rate (CER). In Table 2, the best results for each pair: the test dataset and the speech recognition model are highlighted in bold type.

As can be seen from the results, in almost all cases the accent conversion method improves the recognition of the pretrained models, as indicated by the reduced values of word and character error rates. The accent conversion model improves speech quality by making it more recognizable.

3.4. Subjective quality assessment

Group-based listening tests were conducted with 53 participants from different countries with an English language proficiency level of at least B2 according to the CEFR[4] scale. For this purpose,

[4] CEFR (Common European Framework of Reference) is the system of foreign language proficiency levels used in Europe. https://www.coe.int/en/web/common-european-framework-reference-languages. Accessed May 01, 2024.

Table 2. Results of accent conversion model assessment with speech recognition models. Data after conversion are marked as ‘conv.’

Test dataset	Speech recognition model			
	Conformer	Citrinet	Whisper L. Mult.	Whisper M. En.
WER, %				
ARCTIC	9.57	11.73	16.23	8.91
ARCTIC conv.	**8.78**	**11.55**	**12.69**	**8.68**
Common Voice	**9.07**	25.80	36.89	11.26
Common Voice conv.	9.12	**23.38**	**22.71**	**10.62**
NPTEL2020	29.18	29.88	16.41	15.18
NPTEL2020 conv.	**25.26**	**29.41**	**13.87**	**11.64**
Afrispeech-200	43.2	46.24	37.91	33.61
Afrispeech-200 conv.	**35.19**	**39.49**	**35.56**	**29.96**
CER, %				
ARCTIC	3.73	4.85	10.30	3.98
ARCTIC conv.	**3.52**	**4.68**	**6.06**	**3.92**
Common Voice	**3.75**	8.74	21.41	5.66
Common Voice conv.	3.77	**8.29**	**13.63**	**5.22**
NPTEL2020	16.87	17.70	11.94	10.67
NPTEL2020 conv.	**14.79**	**17.01**	**10.10**	**9.44**
Afrispeech-200	31.52	34.79	24.30	20.04
Afrispeech-200 conv.	**27.86**	**28.92**	**23.15**	**18.88**

each of the participants was given instructions, where within each experiment they were asked to listen to 1 or 2 audio files and give their assessment of compliance with the quality criterion on a five-point scale, where ‘1’ – definitely does not comply, ‘2’ – rather does not comply, ‘3’ – compromise, ‘4’ – rather complies, ‘5’ – exactly complies. The resulting scores were then used to calculate the mean opinion score (MOS) for each experiment. The results are presented in Table 3.

Twenty pairs of audio files from test subsamples of L2 ARCTIC [46] and NPTEL2020 datasets with non-native English accent (Original) were randomly selected as audio samples: Indian, Chinese, Korean, Vietnamese, Spanish, Arabic, and German. Each original audio pair represents a recording of the same speaker. For each selected audio file (40 in total), variants were generated using a simplified accent conversion model (Ablation) and using the proposed model (Proposed). A total of 3 experiments were conducted to evaluate voice naturalness, speaker similarity and absence of foreign accent. In all experiments, at least 3 evaluations were asked for each type of sound sample. The test samples themselves were varied across experiments, eliminating repetition. The sample could be listened to an unlimited number of times before scoring. Thus, each interviewee made a total of 9 to 12 evaluations.

Table 3. Results of subjective quality assessment (MOS with 95% confidence interval)

Examples	Naturalness of voice	Similarity of the speakers	Absence of foreign accent
Original	4.83 ± 0.10	4.91 ± 0.08	2.06 ± 0.18
Ablation	3.38 ± 0.13	3.92 ± 0.15	3.58 ± 0.17
Proposed	4.04 ± 0.16	4.30 ± 0.18	4.11 ± 0.14

When assessing the naturalness of the voice, participants were asked to determine on a five-point scale how natural the speech in the audio example sounds, i.e., whether the listener gets the impression that it is a real live human voice and not a generated or robotized speech. A score of ‘1’ means that the voice is definitely artificial, synthesized using computer-generated methods, and ‘5’ means that the example sounds like speech produced using analog or digital sound recording methods of a real human voice. Interviewees were also advised not to pay attention to the presence or absence of background noise in the recording in order to concentrate on speech evaluation.

In order to conduct an experiment on speaker similarity evaluation, pairs of audio recordings were prepared: Original–Original, Original–Ablation, and Original–Proposed. Meanwhile, the first pair includes recordings from the original data only, which are recordings of the same speaker but uttering different phrases. The other pairs include an original recording of one phrase and a generated version of another phrase by the same speaker. Participants were asked to listen to such pairs of audio recordings and decide whether they were spoken by the same person, i.e., how similar the timbre in one file is to the timbre in the other file. Score ‘1’ is the speech in the audio recordings definitely belongs to different people, ‘5’ is the timbre of the speakers in the audio recordings is identical, belonging to one person. Interviewees were recommended to ignore the L1 and L2 accent properties during the evaluation in order to focus on comparing the overtone coloration of the voice.

To assess the absence of a foreign accent, participants were asked to listen to an English-language audio file and decide how much of a foreign accent they thought the recording contained. English and American accents were assumed to be native L1 and all other accents were assumed to be non-native L2. A score of ‘1’ means that the speech has a pronounced foreign L2 accent, ‘5’ means that the speech is definitely an English-speaking L1 without a foreign accent.

The analysis of the table shows that the highest estimates of voice naturalness and speaker similarity show the original examples, which is obvious, since they are obtained without using speech synthesis methods, and together with the lowest estimate of the absence of foreign accent demonstrates the calibration of the opinions of the experiment participants on real data. Adding the AE/GE model to the overall scheme of the accent conversion model significantly improves the quality of generation, this is demonstrated by the improved results compared to the simplified Ablation model. In all subjective experiments, the proposed model shows a score higher than ‘4’, meaning that, in the opinion of the interviewees, the model rather meets the specified quality criteria.

CONCLUSIONS

The study presents an accent conversion method that converts any L2 speech with a pronounced foreign accent into L1 speech, does not depend on the availability of reference examples and parallel data at the training and generation stages, which greatly simplifies, cheapens and speeds up the process of adapting the system to new accents.

The proposed non-autoregressive model, which does not use recurrent networks in its architecture, features an accelerated training process and real-time accent translation while avoiding the error accumulation effect associated with sequential output generation.

The described method also includes an algorithm for cloning the speech characteristics of the speaker to preserve his or her vocal identity even following accent conversion. This is especially important in situations where emotional coloration, expressiveness, and individual speech features are required. In addition, the method enables real-time modification of voice characteristics such as accent, timbre, and gender-related voice features during the generation process by copying the corresponding characteristics from an audio sample, making it applicable in a wider range of scenarios than previous developments.

The model demonstrates high quality accent conversion while preserving the original timbre, as well as low generation latency acceptable for use in real-time scenarios.

The method can be used to:

1) convert English-speaking speech with a foreign L2 accent into L1 speech without a foreign accent;
2) improve speech quality and, as a consequence, to improve the recognition quality of existing systems;
3) copy and change the speaker's voice characteristics in real time;
4) apply real-time accent conversion in a dialog mode.

The developed applied neural network model demonstrated the ability to work in real-time English language information systems. The results of the study can be applied to the development of voice modification systems, as well as speech recognition and generation systems.

Authors' contribution

All authors equally contributed to the research work.

REFERENCES

1. McMillin D.C. Outsourcing identities: Call centres and cultural transformation in India. *Economic and Political Weekly*. 2006;41(3):235–241.
2. Felps D., Bortfeld H., Gutierrez-Osuna R. Foreign accent conversion in computer assisted pronunciation training. *Speech Communication*. 2009;51(10):920–932. https://doi.org/10.1016/j.specom.2008.11.004
3. Probst K., Ke Y., Eskenazi M. Enhancing foreign language tutors–In search of the golden speaker. *Speech Communication*. 2002;37(3–4):161–173. https://doi.org/10.1016/S0167-6393(01)00009-7
4. Türk O., Arslan L. M. Subband based voice conversion. In: *7th International Conference on Spoken Language Processing, ICSLP2002 – INTERSPEECH 2002*. *Interspeech*. 2002. P. 289–292.
5. Biadsy F., Weis, R.J., Moreno P.J., Kanevsky D., Jia Y. Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation. *Interspeech*. 2019. P. 4115–4119. http://doi.org/10.21437/Interspeech.2019-1789
6. Birner B. *Why Do Some People Have an Accent?* Linguistic Society of America. Washington, DC. 1999. 6 p.
7. Baese-Berk M.M., Morrill T.H. Speaking rate consistency in native and non-native speakers of English. *J. Acoust. Soc. Am*. 2015;138(3):EL223–EL228. https://doi.org/10.1121/1.4929622
8. Piske T., MacKay I.R.A., Flege J.E. Factors affecting degree of foreign accent in an L2: A review. *J. Phonetics.* 2001;29(2):191–215. https://doi.org/10.1006/jpho.2001.0134
9. Munro M.J., Derwing T.M. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*. 1995;45(1):73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x
10. Lev-Ari S., Keysar B. Why don't we believe non-native speakers? The influence of accent on credibility. *J. Exp. Soc. Psychol*. 2010;46(6):1093–1096. https://doi.org/10.1016/j.jesp.2010.05.025
11. Rubin D.L., Smith K.A. Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants. *Int. J. Intercult. Relat*. 1990;14(3):337–353. https://doi.org/10.1016/0147-1767(90)90019-S
12. Nelson Jr. L.R., Signorella M.L., Botti K.G. Accent, gender, and perceived competence. *Hispanic J. Behavior. Sci.* 2016;38(2):166–185. https://doi.org/10.1177/0739986316632319
13. Zhao G., Gutierrez-Osuna R. Using phonetic posteriorgram based frame pairing for segmental accent conversion. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2019;27(10):1649–1660. https://doi.org/10.1109/TASLP.2019.2926754
14. Zhao G., Sonsaat S., Levis J., Chukharev-Hudilainen E., Gutierrez-Osuna R. Accent conversion using phonetic posteriorgrams. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2018. P. 5314–5318. https://doi.org/10.1109/ICASSP.2018.8462258
15. Aryal S., Gutierrez-Osuna R. Can voice conversion be used to reduce non-native accents? In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2014. P. 7879–7883. https://doi.org/10.1109/ICASSP.2014.6855134
16. Ding S., Zhao G., Gutierrez-Osuna R. Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning. *Computer Speech & Language*. 2022;72:101302. https://doi.org/10.1016/j.csl.2021.101302
17. Quamer W., Das A., Levis J., Chukharev-Hudilainen E., Gutierrez-Osuna R. Zero-shot foreign accent conversion without a native reference. *Proc. Interspeech*. 2022. http://doi.org/10.21437/Interspeech.2022-10664
18. Jin M., Serai P., Wu J., Tjandra A., Manohar V., He Q. Voice-preserving zero-shot multiple accent conversion. In: *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2023. P. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10094737
19. Zhou Y., Wu Z., Zhang M., Tian X., Li H. TTS-guided training for accent conversion without parallel data. *IEEE Signal Proc. Lett*. 2023;30:533–537. https://doi.org/10.1109/lsp.2023.3270079
20. Zhao G., Ding S., Gutierrez-Osuna R. Converting foreign accent speech without a reference. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2021;29:2367–2381. https://doi.org/10.1109/TASLP.2021.3060813
21. Liu S., Wang D., Cao Y., Sun L., Wu X., Kang S., Wu Z., Liu X., Su D., Yu D., Meng H. End-to-end accent conversion without using native utterances. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2020. P. 6289–6293.

22. Zhou X., Zhang M., Zhou Y., Wu Z., Li H. Accented text-to-speech synthesis with limited data. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2024;32:1699–1711. https://doi.org/10.1109/TASLP.2024.3363414
23. Pinget A.F., Bosker H.R., Quené H., De Jong, N.H. Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing*. 2014;31(3):349–365. https://doi.org/10.1177/0265532214526177
24. Barkhudarova E.L. Methodological Problems in Analyzing Foreign Accents in Russian Speech. *Vestnik Moskovskogo universiteta. Seriya 9. Filologiya = Lomonosov Philology J.* 2012;6:57–70 (in Russ.).
25. Arik S., Chen J., Peng K., Ping W., Zhou Y. Neural voice cloning with a few samples. *Advances in Neural Information Processing Systems* (*NeurIPS 2018*). 2018;31.
26. Cohen D. Issues in transnet packetized voice communication. In: *Proceedings of the fifth Symposium on Data Communications* (*SIGCOMM'77*). 1977. P. 6.10–6.13. https://doi.org/10.1145/800103.803349
27. Liang Y.J., Farber N., Girod B. Adaptive playout scheduling and loss concealment for voice communication over IP networks. *IEEE Trans. Multimedia*. 2003;5(4):532–543. https://doi.org/10.1109/TMM.2003.819095
28. Matzinger T., Pleyer M., Żywiczyński P. Pause Length and Differences in Cognitive State Attribution in Native and Non-Native Speakers. *Languages*. 2023;8(1):26. http://doi.org/10.3390/languages8010026
29. Medsker L.R., Jain L. (Eds.). *Recurrent Neural Networks. Design and Applications*. Boca Raton: CRC Press; 2001. 416 p.
30. Li J., Lavrukhin V., Ginsburg B., Leary R., Kuchaiev O., Cohen J.M., Nguyen H., Gadde R.T. Jasper: An End-to-End Convolutional Neural Acoustic Model. *Interspeech 2019*. 2019. https://doi.org/10.21437/interspeech.2019-1819
31. Dawalatabad N., Ravanelli M., Grondin F., Thienpondt J., Desplanques B., Na H. *ECAPA-TDNN Embeddings for Speaker Diarization.* arXiv preprint arXiv:2104.01466. 2021. https://doi.org/10.48550/arXiv.2104.01466
32. Ravanelli M., Bengio Y. Speaker recognition from raw waveform with SincNet. In: *2018 IEEE Spoken Language Technology Workshop* (*SLT*). IEEE; 2018. P. 1021–1028. https://doi.org/10.1109/SLT.2018.8639585
33. Snyder D., Garcia-Romero D., Sell G., Povey D., Khudanpur S. X-vectors: Robust DNN embeddings for speaker recognition. In: *2018 IEEE international Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2018. P. 5329–5333. http://doi.org/10.1109/ICASSP.2018.8461375
34. Deng J., Guo J., Xue N., Zafeiriou S. Arcface: Additive angular margin loss for deep face recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE; 2019. P. 4690–4699. https://doi.org/10.1109/CVPR.2019.00482
35. Gulati A., Qin J., Chiu C., Parmar N., Zhang Y., Yu J., Han W., Wang S., Zhang Z., Wu Y., Pang R. Conformer: Convolution-augmented trans-former for speech recognition. *Proc. Interspeech 2020*. 2020. P. 5036–5040. https://doi.org/10.21437/interspeech.2020-3015
36. Glorot X., Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research*. 2010. P. 249–256. URL: http://proceedings.mlr.press/v9/glorot10a.html
37. Gu J., Wang Z., Kuen J., Ma L., Shahroudy A., Shuai B., Liu T., Wang X., Wang G., Cai J., Chen T. Recent advances in convolutional neural networks. *Pattern Recognition*. 2018;77:354–377. https://doi.org/10.1016/j.patcog.2017.10.013
38. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need. *Advances in Neural Information Processing Systems*. 2017;30:5999–6009. https://doi.org/10.48550/arXiv.1706.03762
39. Ren Y., Ruan Y., Tan X., Qin T., Zhao S., Zhao Z., Liu T.Y. FastSpeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*. 2019;32. https://doi.org/10.48550/arXiv.1905.09263
40. Graves A., Fernández S., Gomez F., Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: *Proceedings of the 23rd International Conference on Machine Learning*. 2006. P. 369–376. https://doi.org/10.1145/1143844.1143891
41. Ghahremani P., BabaAli B., Povey D., Riedhammer K., Trmal J., Khudanpur S. A pitch extraction algorithm tuned for automatic speech recognition. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2014. P. 2494–2498. http://doi.org/10.1109/ICASSP.2014.6854049
42. Gerhard D. *Pitch Extraction and Fundamental Frequency: History and Current Techniques*. Masters Thesis. Regina, SK, Canada: Department of Computer Science, University of Regina; 2003. 23 p.
43. Łańcucki A. Fastpitch: Parallel text-to-speech with pitch prediction. In: *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2021. P. 6588–6592. https://doi.org/10.1109/ICASSP39728.2021.9413889
44. Kong J., Kim J., Bae J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*. 2020;33:17022–17033. http://doi.org/10.48550/arXiv.2010.05646
45. Kominek J., Black A.W. The CMU Arctic speech databases. In: *Fifth ISCA Workshop on Speech Synthesis*. 2004. P. 223–224.
46. Zhao G., Sonsaat S., Silpachai A., Lucic I., Chukharev-Hudilainen E., Levis J., Gutierrez-Osuna R. L2-ARCTIC: A Non-native English Speech Corpus. *Interspeech 2018*. 2018. P. 2783–2787. http://doi.org/10.21437/Interspeech.2018-1110
47. Weinberger S.H., Kunath S.A. The Speech Accent Archive: towards a typology of English accents. In: *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. Brill; 2011. P. 265–281. https://doi.org/10.1163/9789401206884_014
48. Ardila R., Branson M., Davis K., Kohler M., Meyer J., Henretty M., Henretty M., Morais R., Saunders L., Tyers F., Weber G. Common Voice: A Massively-Multilingual Speech Corpus. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020. P. 4218–4222. https://doi.org/10.48550/arXiv.1912.06670

49. Nagrani A., Chung J.S., Zisserman A. Voxceleb: a large-scale speaker identification dataset. *Interspeech 2017*. 2017. http://doi.org/10.21437/Interspeech.2017-950
50. Chung J., Nagrani A., Zisserman A. VoxCeleb2: Deep speaker recognition. *Interspeech 2018*. 2018. http://doi.org/10.21437/Interspeech.2018-1929
51. Panayotov V., Chen G., Povey D., Khudanpur S. Librispeech: an ASR corpus based on public domain audio books. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2015. P. 5206–5210. http://doi.org/10.1109/ICASSP.2015.7178964
52. Veaux C., Yamagishi J., MacDonald K. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. *University of Edinburgh. The Center for Speech Technology Research* (*CSTR*). 2017. https://doi.org/10.7488/ds/2645
53. Chen G., Chai S., Wang G., Du J., Zhang W., Weng C., Su D., Povey D., Trmal J., Zhang J., Jin M., Khudanpur S., Watanabe S., Zhao S., Zou W., Li X., Yao X., Wang Y., Wang Y., You Z., Yan Z. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In: *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021*. International Speech Communication Association; 2021. P. 4376–4380. https://doi.org/10.21437/Interspeech.2021-1965
54. Bakhturina E., Zhang Y., Ginsburg B. Shallow Fusion of Weighted Finite-State Transducer and Language Model for Text Normalization. *Proc. Interspeech 2022*. 2022. http://doi.org/10.48550/arXiv.2203.15917
55. Kudo T., Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2018. P. 66–71. https://doi.org/10.48550/arXiv.1808.06226
56. Koizumi Y., Zen H., Karita S., Ding Y., Yatabe K., Morioka N., Bacchiani M., Zhang Y., Han W., Bapna A. Libritts-r: A Restored Multi-Speaker Text-to-Speech Corpus. *arXiv preprint arXiv:2305.18802*. 2023. https://doi.org/10.48550/arXiv.2305.18802
57. Paszke A., Gross S., Massa F., Lerer A., Bradbury J., Chanan G., Killeen T., Lin Z., Gimelshein N., Antiga L., Desmaison A., Kopf A., Yang E., DeVito Z., Raison M., Tejani A., Chilamkurthy S., Steiner B., Fang L., Bai J., Chintala S. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*. 2019;32:8024–8035.
58. Kuchaiev O., Li J., Nguyen H., Hrinchuk O., Leary R., Ginsburg B., Kriman S., Beliaev S., Lavrukhin V., Cook J., Castonguay P., Popova M., Huang J., Cohen J. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*. 2019. https://doi.org/10.48550/arXiv.1909.09577
59. Bredin H., Yin R., Coria J.M., Gelly G., Korshunov P., Lavechin M., Fustes D., Titeux H., Bouaziz W., Gill M.P. Pyannote. Audio: neural building blocks for speaker diarization. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE; 2020. P. 7124–7128. https://doi.org/10.1109/ICASSP40776.2020.9052974
60. Olatunji T., Afonja T., Yadavalli A., Emezue C.C., Singh S., Dossou B.F., Osuchukwu J., Osei S., Tonja A.L., Etori N., Mbataku C. AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR. *Transactions of the Association for Computational Linguistics*. 2023;11:1669–1685. https://doi.org/10.1162/tacl_a_00627
61. Majumdar S., Balam J., Hrinchuk O., Lavrukhin V., Noroozi V., Ginsburg B. Citrinet: Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition. *arXiv preprint arXiv:2104.01721*. 2021. http://doi.org/10.48550/arXiv.2104.01721
62. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR 202. 2023. P. 28492–28518. http://doi.org/10.48550/arXiv.2212.04356

About the Authors

Vladimir A. Nechaev, Teacher-Researcher, Ivanovo State Power Engineering University (34, Rabfakovskaya ul., Ivanovo, 153003 Russia). E-mail: nechaev@gapps.ispu.ru. RSCI SPIN-code 7002-3878, https://orcid.org/0009-0007-1449-3968

Sergey V. Kosyakov, Dr. Sci. (Eng.), Professor, Head of the Department of Computer Systems Software, Ivanovo State Power Engineering University (34, Rabfakovskaya ul., Ivanovo, 153003 Russia). E-mail: ksv@ispu.ru. Scopus Author ID 6507182528, ResearcherID H-5686-2018, RSCI SPIN-code 1371-9929, https://orcid.org/0000-0003-0231-0750

Об авторах

Нечаев Владимир Алексеевич, преподаватель-исследователь, ФГБОУ ВО «Ивановский государственный энергетический университет имени В.И. Ленина» (153003, Россия, Иваново, ул. Рабфаковская, д. 34). E-mail: nechaev@gapps.ispu.ru. SPIN-код РИНЦ 7002-3878, https://orcid.org/0009-0007-1449-3968

Косяков Сергей Витальевич, д.т.н., профессор, заведующий кафедрой программного обеспечения компьютерных систем, ФГБОУ ВО «Ивановский государственный энергетический университет имени В.И. Ленина» (153003, Россия, Иваново, ул. Рабфаковская, д. 34). E-mail: ksv@ispu.ru. Scopus Author ID 6507182528, ResearcherID H-5686-2018, SPIN-код РИНЦ 1371-9929, https://orcid.org/0000-0003-0231-0750

Translated from Russian into English by L. Bychkova
Edited for English language and spelling by Thomas A. Beavitt

Information systems. Computer sciences. Issues of information security

Информационные системы. Информатика. Проблемы информационной безопасности

UDC 004.8
https://doi.org/10.32362/2500-316X-2025-13-3-21-43
EDN QKUGFZ

REVIEW ARTICLE

Knowledge injection methods in question answering

Daniil V. Radyush @

ITMO University, Saint Petersburg, 197101 Russia
@ *Corresponding author, e-mail: daniil.radyush@gmail.com*

• **Submitted:** 26.06.2024 • **Revised:** 13.02.2025 • **Accepted:** 27.03.2025

Abstract

Objectives. Despite the recent success of large language models, which are now capable of solving a wide range of tasks, a number of practical issues remain unsolved. For example, users of systems providing question answering (QA) services may experience a lack of commonsense knowledge and reasoning proficiency. The present work considers knowledge injection methods as a means of providing functional enhancements to large language models by providing necessary facts and patterns from external sources.

Methods. Knowledge injection methods leveraged in relevant QA systems are classified, analyzed, and compared. Self-supervised learning, fine-tuning, attention mechanism and interaction tokens for supporting information injection are considered along with auxiliary approaches for emphasizing the most relevant facts.

Results. The reviewed QA systems explicitly show the accuracy increase on the CommonsenseQA benchmark compared to pretrained language model baseline due to knowledge injection methods exploitation. At the same time, in general the higher results are related to knowledge injection methods based on language models and attention mechanism.

Conclusions. The presented systematic review of existing external knowledge injection methods for QA systems confirms the continuing validity of this research direction. Such methods are not only capable of increasing the accuracy of QA systems but also mitigating issues with interpretability and factual obsolescence in pretrained models. Further investigations will be carried out to improve and optimize different aspects of the current approaches and develop conceptually novel ideas.

Keywords: deep learning, nature language processing, question answering system, knowledge base, graph neural networks, knowledge injection

For citation: Radyush D.V. Knowledge injection methods in question answering. *Russian Technological Journal.* 2025;13(3):21–43. https://doi.org/10.32362/2500-316X-2025-13-3-21-43, https://www.elibrary.ru/QKUGFZ

Financial disclosure: The author has no financial or proprietary interest in any material or method mentioned.

The author declares no conflicts of interest.

© D.V. Radyush, 2025

ОБЗОР

Методы интеграции знаний для разработки вопросно-ответных систем

Д.В. Радюш @

Национальный исследовательский университет ИТМО, Санкт-Петербург, 197101 Россия
@ Автор для переписки, e-mail: daniil.radyush@gmail.com

• **Поступила:** 26.06.2024 • **Доработана:** 13.02.2025 • **Принята к опубликованию:** 27.03.2025

Резюме

Цели. Несмотря на наблюдаемые в последние несколько лет успехи больших языковых моделей, которые способны решать широкий перечень задач, ряд практических проблем остается не до конца решенным. В контексте построения вопросно-ответных систем к таким проблемам можно отнести использование общих знаний и учет причинно-следственных связей. Целью статьи является рассмотрение методов интеграции знаний, которые способны усовершенствовать функционирование больших языковых моделей путем предоставления необходимых сведений и закономерностей из внешних источников.

Методы. В работе осуществляются классификация, анализ и сопоставление методов интеграции знаний, используемых в актуальных реализациях вопросно-ответных систем. В частности, рассматривается вовлечение вспомогательных сведений через самообучение, дообучение, механизм внимания и использование токенов взаимодействия, а также описываются соответствующие вспомогательные подходы для акцентирования наиболее релевантных сведений.

Результаты. Рассмотренные в обзоре вопросно-ответные системы непосредственно демонстрируют возрастание точности относительно базового решения на основе предобученной языковой модели за счет использования методов интеграции знаний на примере бенчмарка CommonsenseQA. При этом в целом более высокие результаты показывают методы интеграции знаний, основанные на использовании языковых моделей и механизма внимания.

Выводы. Представленный систематический обзор существующих методов интеграции знаний из внешних источников в работу вопросно-ответных систем фактически подтверждает эффективность и перспективность этого направления исследований. Данные методы демонстрируют не только возможность увеличить точность вопросно-ответных систем, но и в некоторой степени сгладить проблемы, связанные с интерпретируемостью результатов и устареванием знаний в предобученных моделях. Последующие изыскания способны как улучшить и оптимизировать отдельные аспекты существующих подходов, так и выработать концептуально новые.

Ключевые слова: глубокое обучение, обработка естественного языка, вопросно-ответная система, база знаний, графовые нейронные сети, интеграция знаний

Для цитирования: Радюш Д.В. Методы интеграции знаний для разработки вопросно-ответных систем. *Russian Technological Journal*. 2025;13(3):21–43. https://doi.org/10.32362/2500-316X-2025-13-3-21-43, https://www.elibrary.ru/QKUGFZ

Прозрачность финансовой деятельности: Автор не имеет финансовой заинтересованности в представленных материалах или методах.

Автор заявляет об отсутствии конфликта интересов.

INTRODUCTION

The development of question answering (QA) systems in recent years has been significantly influenced by the emergence and subsequent improvement of pretrained language models [1]. The effectiveness of such models is based on the processing of large text corpora containing heterogeneous information, which makes it possible to capture both certain linguistic regularities and specific facts in the model weights [2]. Nevertheless, due to the peculiarities of natural languages, a significant amount of relevant information about the surrounding world may not always presented in the text in an explicit form, making it difficult to identify such information by language models at the training stage. In the first place, such information concerns various kinds of social interactions, including their psychological aspects, but also basic physical principles, which are learned by human beings at an early age. Examples of the latter include understanding the need to take care when crossing the road or cool down excessively hot food before eating it.

In order to compensate for this disadvantage, different kinds of knowledge sources can be used, in which such data will be recorded in an unambiguous form. For example, Cyc[1], which set out to collect general ideas about the world around us, can be considered as one of the first examples of a combined ontology and knowledge base. The main idea consists in the information record in the form of logical rules, which corresponded to the main direction of development of applications in the field of artificial intelligence of that time. A number of similar sources have been developed to date, albeit having somewhat different approaches to knowledge description, such as ATOMIC [3] and ConceptNet[2].

From a formal point of view, several arguments in favor of using external knowledge sources in the development of QA systems can be distinguished. Here, the primary motivation is directly to obtain more accurate and satisfying results for the user. By providing additional context to the query, a model can be expected to be able to answer a number of questions for which the internal representations of pretrained language models may not be sufficient. On the one hand, such questions include those where some causal relations are omitted, while, on the other, there is uncertainty in terms of identifying the semantics of some words due to their polysemy and insufficient context. This is largely determined by the limitations identified in the analysis of the application of transformers, which tend to rely only on the superficial and statistically most likely meanings of individual words [4], while for logical inference they rely heavily on heuristics learned from the training sample [5].

Even language models with a large number of weights, which demonstrate high results on many benchmarks, can not only make mistakes, but sometimes produce answers that have no relation to reality, which are popularly known as hallucinations [6]. In this regard, there is even a separate research area dedicated to methods of extracting query-relevant information and including it into the input data to improve the quality of answers [7] and using retrieval-augmented generation to reduce the number of hallucinations [8].

External knowledge sources can be used to reduce the requirements for the necessary computational resources to utilize pretrained language models. In particular, the purposeful involvement of additional information may enable the use of models having fewer weights while maintaining system accuracy at a comparable level [9]. This approach can be used to simplify QA system exploitation, as well as offsetting the cost of extracting and processing auxiliary data.

No less important is the use of knowledge bases from the position of explainable artificial intelligence (XAI). Due to the structured nature of knowledge bases, information extracted from them can provide a sequence of logical reasoning to the user to serve as a justification of the system's result. This property can be extremely important from the point of view of practical application since it is often the lack of interpretability of the results obtained with the help of neural networks that restricts their application in areas where there is a high risk of error and corresponding potential damage to society. In general, in order to effectively evaluate the adequacy of a system, it is desirable to have the most complete understanding of its operation.

Finally, the problem of updating the facts captured in the weights of language models is significant. Training of such models is typically performed on specific data sets and takes quite a long period of time. At the same time, a huge number of events occur every day, which leads to the change of a part of knowledge and the appearance of new facts. One way to solve this problem is to extract such information from external databases.

In this regard, it is quite relevant to consider how to use auxiliary general information to solve specific problems, such as the development of QA systems. In particular, the successful operation of the system requires that the information obtained is sufficient but not redundant, as otherwise it may impede its functioning and degrade the results. Also of equal importance is the way in which the additional knowledge is processed, as this will largely determine the effect of its use in the system. Thus, since the procedure of knowledge injection can be influenced by a substantial number of aspects, this paper presents an attempt to systematically analyze and compare existing

[1] Cycorp. https://cyc.com. Accessed December 01, 2024.

[2] ConceptNet. An open, multilingual knowledge graph. https://conceptnet.io. Accessed December 01, 2024.

approaches in order to draw a complete picture of the relevant ideas.

KNOWLEDGE BASES

In general, several directions can be distinguished in the field of QA system development depending on the methods used to provide additional data. For example, open domain QA implies the absence of a specialized knowledge base and is focused on the use of information from general-purpose sources. For this purpose, Wikipedia[3] is typically used and thus, due to its considerable volume and structural heterogeneity of content, the focus shifts significantly towards methods of searching for relevant information.

When using more narrowly focused queries to search for answers to more specialized questions, closed domain QA approaches include those complicated by the need to perform logical inference and take specific information into account. In this regard, specialized bases of structured knowledge, for example, knowledge graphs, can act as external knowledge sources that simplifies to a certain extent information retrieval, as well as implementation of logical inference and auxiliary operations. A relevant example is the knowledge graph DBpedia[4].

In the context of developing approaches to knowledge injection in the field of QA systems, structured knowledge bases are of primary interest. To some extent, this can be justified by the current state of affairs in this area. In particular, the emergence and subsequent development of pretrained language models has significantly reduced the need for contextual requirements to be provided to the query. As a result, some models after fine-tuning are now able to show results comparable to those of humans. Consequently, the main interest shifts towards analyzing the cases in which humans outperform existing QA systems. As a rule, such queries are those requiring out-of-context general ideas about the structure of the surrounding world, as well as analyzing cause-and-effect relations between individual facts.

In such circumstances, structured common knowledge bases can be particularly useful. In the first place, they directly provide the system with missing facts, which can be extracted taking into account existing relationships among themselves and together with other related information. Moreover, the structured nature of the knowledge greatly simplifies its machine processing and hence its use in practice. Thus, it seems possible to solve, to some extent, simultaneously the problems related to the class of queries that can be considered challenging for existing QA systems.

While Wikipedia can still be of use as an external source of additional data, due to the lack of systematization and large redundancy of information, the Wikidata[5] structured knowledge base which is based on data from Wikipedia has become of increasing relevance. The Wikidata graph consists of more than 100 mln entries describing elements of human knowledge in some way. Since each element of the graph corresponds to a certain set of properties that characterize it and establish its relationships with other elements, the content of Wikidata can be represented as for other knowledge graphs as a set of so-called subject–predicate–object triplets for which the object is a set of specific property values or a reference to another entity.

The ConceptNet knowledge base is also frequently used as a knowledge source in the context of building QA systems. This knowledge base, in addition to unique general information, partially includes information from other relatively frequently used sources such as the previously mentioned Cyc and DBpedia. Within ConceptNet, words and phrases are grouped based on several dozen relations. Comparable to the Wikidata resource discussed earlier, ConceptNet contains over 30 mln entries, although one must keep in mind that a significant portion of this value is due to the presence of effectively duplicate entries due to the existence of counterparts in another language, single-rooted words, and symmetric relationships. In addition, a slightly greater emphasis in ConceptNet is placed on linguistic properties, for example by capturing synonyms, antonyms and etymologically related words for a word. Finally, a feature of ConceptNet is the existence of weights for each relationship between elements, which heuristically reflects the degree of probability or importance of a given relationship.

Among the relatively recent general knowledge bases, ATOMIC, which contains more than 1 mln elements, can also be emphasized. The peculiarity of ATOMIC is the reflection of information in the form of abstract events and their results, which can be used emphasize complex cause-and-effect relations existing in the surrounding world. In particular, for example, based on some event in ATOMIC, it is possible to identify any of its consequences, as well as the intention, desire, or other characteristic of one of the participants, which can provide models with potentially missing knowledge about social interactions.

Table 1 presents examples of information that can be extracted from the knowledge bases discussed above. In general, they are somewhat similar, except for the more specific purpose of the ATOMIC database.

[3] https://www.wikipedia.org/. Accessed December 01, 2024.

[4] The DBpedia Knowledge Base. https://www.dbpedia.org. Accessed December 01, 2024.

[5] Wikidata. The free knowledge base. https://www.wikidata.org. Accessed December 01, 2024.

Table 1. Examples of information extracted from the knowledge bases

Knowledge base	Data example
DBpedia	DBpedia subject SemanticWeb
Wikidata	Wikidata uses semantic technology
ConceptNet	ConceptNet motivated by goal let computers understand what people already know
ATOMIC	Person X pays Person Y a compliment. Person X wanted to be nice

KNOWLEDGE INJECTION METHODS

A classification of knowledge injection methods based on the analysis of current research on the topic is presented in Fig. 1. Accordingly, the main ideas of knowledge injection methods can be considered in the context of developing QA systems with the corresponding examples and taking into account the peculiarities of specific selected classes of methods. The main place in this classification is occupied by the division of knowledge injection methods according to the use of knowledge bases, which is understood as processing the information directly when inferencing answers to queries and thus excludes cases of involving knowledge bases in the process of pretraining models. In turn, both language and graph models can be used to extract features from knowledge base data.

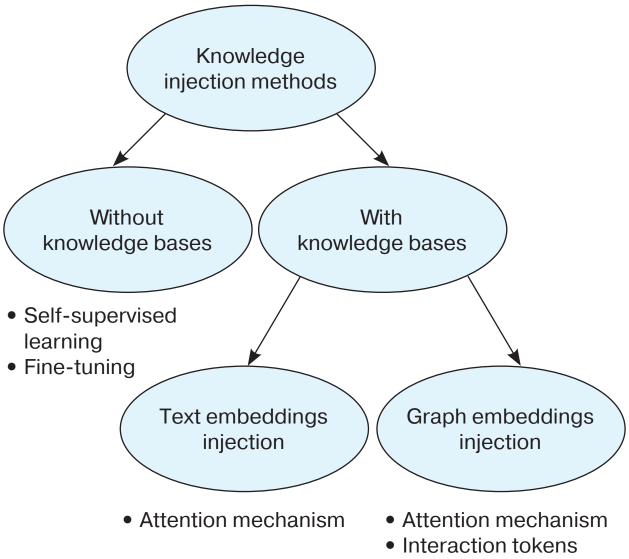


Fig. 1. Classification of knowledge injection methods

METHODOLOGICAL BASIS

Despite the differences in the approaches used for knowledge injection, it is possible to identify a common methodological basis in the QA systems discussed below. In particular, this applies both to the problem formulation itself and the supporting methods used.

The use of additional context in the system creates a certain specificity in terms of operation. In this regard, such auxiliary stages as retrieval of relevant data to the query also begin to acquire significant importance. In general, this stage implies determination of some number of entities n: $(q_1, \ldots, q_n)$ in the received query. For this purpose, classical methods from the field of natural language processing, such as lemmatization and part-of-speech tagging, continue to be mostly used in practice. The subsequent part of the process may vary depending on the specific task.

Many works on knowledge injection in QA systems assume that a question has answer choices. Accordingly, the goal of the system is to estimate the probability of each answer and select the most probable one. This allows us to significantly simplify and unify the construction and evaluation of systems. Therefore, in such cases, we will assume that the identified n entities correspond to m similarly extracted entities from the answer choices: $(a_1, \ldots, a_m)$.

The next step involves some knowledge base, which can be formalized as $G = (V, E)$, where V is the set of entities in the knowledge base, and $E \subseteq V \times R \times V$ is the set of triplets of the knowledge base of the entity–relation–entity kind. In practice, the established form of representation of such knowledge bases is a graph. Based on this, it is possible to construct a set of paths between entities defined in the context of the used knowledge base of the following form:

$$p = ((q_i, r_l, v_l), (v_l, r_{l+1}, v_{l+1}), \ldots, (v_{k-1}, r_k, a_j)),$$

where $i \in (1, \ldots, n)$, $j \in (1, \ldots, m)$; k is the path length in the graph; $l \in (1, \ldots, k)$; q_i is the ith entity from the query; a_j is the jth entity from the answer; v_l and r_l are the lth entity and relation in the graph, respectively.

Subsequently, a knowledge base subgraph or set of paths is used as additional context to determine the most likely answer.

One of the main problems in this setting is to determine the most relevant information in relation to the query. A possible tool for solving this problem is the Attention mechanism [10], which allows us to calculate the so-called Attention Weights quantitatively evaluating the degree of importance of this or that information from the context. Formally, the attention mechanism is defined by the expression:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\times\mathbf{K}^{\text{T}}}{\sqrt{d_{\text{model}}}}\right)\cdot\mathbf{V} = \quad (1)$$

$$= \textbf{Attention weights}\cdot\mathbf{V},$$

where $\mathbf{Q}$ = Query = $\mathbf{X}\times\mathbf{W}_{\text{q}}$; $\mathbf{K}$ = Key = $\mathbf{X}\times\mathbf{W}_{\text{k}}$; $\mathbf{V}$ = Value = $\mathbf{X}\times\mathbf{W}_{\text{v}}$; $\mathbf{X}\in\mathbb{R}^{N\times d_{\text{model}}}$ is the embedding[6] of the input data; $\mathbf{W}_{\text{q}}\in\mathbb{R}^{d_{\text{model}}\times d_k}$, $\mathbf{W}_{\text{k}}\in\mathbb{R}^{d_{\text{model}}\times d_k}$, $\mathbf{W}_{\text{v}}\in\mathbb{R}^{d_{\text{model}}\times d_v}$, N is the number of vectors in the input data; d_{model}, d_k, d_v are the dimensions of the embedding in the model and matrices $\mathbf{K}$ and $\mathbf{V}$, and $\text{softmax}(X_i) = \dfrac{\exp(X_i)}{\sum_{j=1}^{N}\exp(X_j)}$.

Thus, the attention weights, when multiplied by the embedding of the context, can adjust the influence of its individual elements on the result. In practice, it is common to use several groups of different matrices (so-called *heads*) to take into account different aspects of the data within the mechanism; the ensuing result of their application is concatenated and projected into the desired dimension using one more matrix, which is called Multi-Head Attention:

$$\text{Multi-Head Attention} =$$
$$= \text{Concatenation}(\mathit{Attention}_1, \ldots, \mathit{Attention}_z)\times W_{\text{o}}, \quad (2)$$

where $\mathit{Attention}_i$ is the ith result of the Attention block, $W_{\text{o}}\in\mathbb{R}^{zd_v\times d_{\text{model}}}$, while z is the number of the attention heads.

The attention mechanism, which plays a major role in many deep learning models, is widely used in developing approaches for knowledge injection. A feedforward neural network is also often used in conjunction with multi-head attention, which together form the main part of the model called a transformer. A multilayer perceptron is a specific implementation of feedforward neural network; from a practical point of view, the role of this component of the transformer is considered in the context of storing and retrieving patterns learned in the process of the training.

METHODS OF KNOWLEDGE INJECTION WITHOUT KNOWLEDGE BASES

The involvement of auxiliary knowledge does not necessarily imply the use of specific knowledge bases. For example, similar examples with an indication of the correct answer can be used as information to help obtain a more accurate answer to a query. For example, [11] and [12] demonstrate the positive effect of adding queries from the training sample to the input data based on the similarity of their embeddings to the embedding of the original query.

Another type of approach is based on the idea of directly accessing the information learned in the process of the model pretraining; retrieved depending on the query, it is used as additional input data. For this purpose, [13] proposes to ask the pretrained model clarifying questions using templates, and use the answers as useful context. A similar approach in [14] also involves the exploitation of auxiliary data generated for a query in the QA system. Specially trained for knowledge generation models as described in [15] generate structured information in the format of knowledge base paths. Thus, the approaches discussed above are based on the idea of providing additional information as input, for which pretrained and other auxiliary models can be used, while fine-tuning of the language models may not be required.

Quantitatively, the most extensive group of approaches comes from the concept of pretraining models. Many experimental results support the idea that models with a large number of weights, trained on as much diverse information as possible, are able to show better results when they are subsequently adapted to specific conditions [16]. This concept relies heavily on the self-supervised learning methodology, which enables the extraction of representations from text corpora without the need of labels. To accomplish this, special tasks are developed according to the model training requirements. In particular, two such tasks were used in the development of the bidirectional encoder representations from transformers (BERT) language model [1]. The first task is the prediction of words in a sentence masked by a special token. For this purpose, some tokens from a sentence are selected with a probability of 15%, then 80% of these tokens are masked, 10% are replaced by a random token, while the remaining 10% are left unchanged. Cross-entropy can be used as a measure of error for the model's masked token prediction:

$$\text{Cross-entropy} = -\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i\times\log(\overline{\mathbf{y}}_i), \quad (3)$$

[6] Embedding means a vector representation.

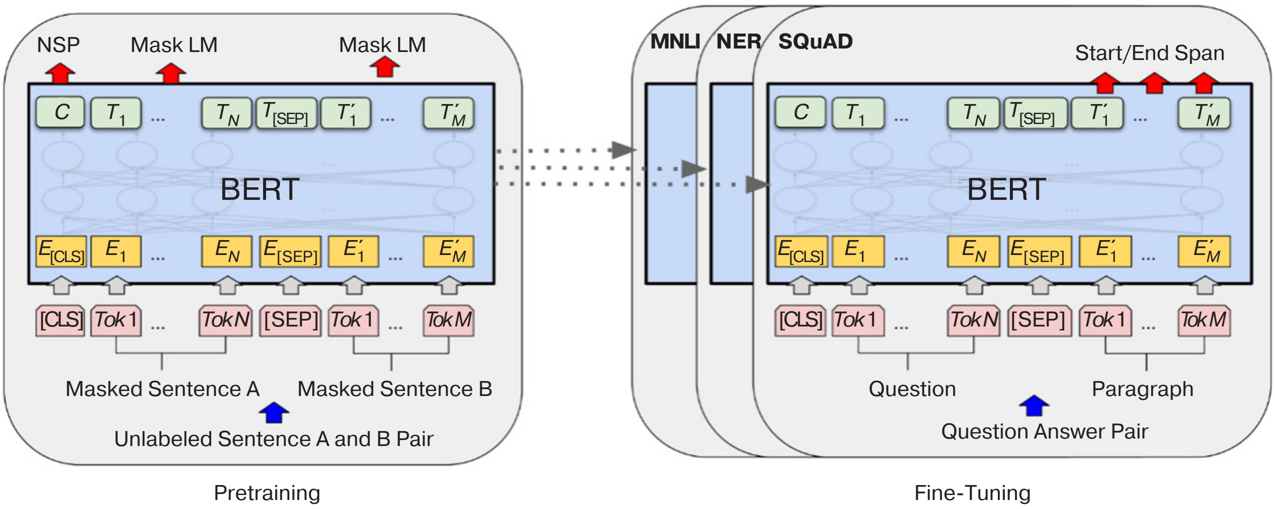


Fig. 2. BERT model training scheme [1]

where N is the total number of examples; $\mathbf{y}_i$ is one-hot-vector[7] encoding the correct answer for the ith example; $\bar{\mathbf{y}}_i$ is the model prediction vector for the ith example denoting the probability of matching each possible answer choice within the problem.

The second task concerns determining the correct order of two sentences in a text. This is realized through adding a special token [CLS] to the input data during training, representing information from the entire sentence, and is reduced to a binary classification task. The goal of this classification is to determine whether some sentence B is a continuation for the sentence A based on their resulting embeddings in the output of the model for the [CLS] tokens. In this learning framework, in 50% of cases, B is a random sentence, while in the other 50% of cases, B is a correct continuation. Cross-entropy can also function as the measure of loss for the task. The total model loss during training is considered as the sum of losses for each task. The overall training scheme of the BERT model is shown in Fig. 2:

In the pretraining phase, some tokens of unlabeled sentence A and B pair are masked, after which the embeddings ($E_{[CLS]}$, E_1, …, E_N, $E_{[SEP]}$, and $E_1{}'$, …, $E_M{}'$) of the tockens (*Tok* 1, …, *Tok N* and *Tok* 1, …, *Tok M*) of the masked sentence A and masked sentence B with addition of generalization and separation tokens ([CLS] and [SEP]) are fed to the input of the BERT transformer. The resulting final embeddings (C, T_1, …, T_N, T_{SEP}, $T_1{}'$, …, $T_M{}'$) are used to predict masked tokens and sentence order. In the Fine-Tuning phase, the format of the input data and the predicted data changes depending on the task (MNLI[8], NER[9], SQuAD[10]). In the case of QA SQuAD dataset, the input is a Question and the corresponding Paragraph, and the output is a predicted position in the context of the correct answer (Start/End Span).

Subsequently, this methodology has been modified and adapted in the context of pretraining of other language models. In the context of QA systems, many approaches have been developed based on modifying and extending BERT self-supervised learning tasks or replacing them with others. In general, when building these kinds of models, it is most common to modify the masking procedure by imposing constraints on what should be masked in a sentence and changing the masking parameters during training.

One of the first and most significant developments in this direction was the Enhanced Language Representation with Informative Entities (ERNIE) model [17], the scheme of which is shown in Fig. 3. Its main idea is that if one also pretrains to predict the masked named entities identified in the text based on the knowledge base as an additional task for self-supervised learning, it can improve the model's language understanding as well as contextualize its certain knowledge about the world. Specifically, for this purpose, for text tokens, the corresponding named entity is replaced by a random entity in 5% of the cases, and in 15% of the cases, the entity is masked and should be predicted from the text tokens. In addition, the paper introduces an interaction

[7] One-hot vector is a binary vector in which only one element has the value 1, and remaining elements are equal to 0.

[8] Multi-Genre Natural Language Inference—dataset for the Natural Language Inference task—establishes a logical relationship between the text fragments.

[9] Named Entity Recognition is the task of recognizing named entities in the text.

[10] Stanford Question Answering Dataset is a QA dataset, which implies automatic answers to questions in natural language.

mechanism between the embeddings of entities and the corresponding text tokens, which can bring additional information to both types of embeddings, thereby increasing the accuracy of predicting the correct tokens. To this end, an intermediate embedding is introduced that combines information at the level of tokens and named entities, due to which the initial embeddings of tokens and entities are subsequently updated, which is defined by the expressions:

$$\begin{gathered}\mathbf{h}_j = \sigma(\widetilde{\mathbf{W}}_\mathrm{t} \widetilde{\mathbf{w}}_j + \widetilde{\mathbf{W}}_\mathrm{e} \tilde{\mathbf{e}}_k + \tilde{\mathbf{b}}),\\ \mathbf{w}_j = \sigma(\mathbf{W}_\mathrm{t} \mathbf{h}_j + \mathbf{b}_\mathrm{t}),\\ \mathbf{e}_k = \sigma(\mathbf{W}_\mathrm{e} \mathbf{h}_j + \mathbf{b}_\mathrm{e}),\end{gathered} \tag{4}$$

where $\mathbf{h}_j$ is the aggregate embedding of the token number j, σ is a given nonlinear activation function, $\widetilde{\mathbf{w}}_j$ and $\mathbf{w}_j$ are embeddings of the token j before and after knowledge injection, $\tilde{\mathbf{e}}_k$ and $\mathbf{e}_k$ are the embeddings of the named entity k matching the token j before and after knowledge injection, $\widetilde{\mathbf{W}}$, $\mathbf{W}$, $\tilde{\mathbf{b}}$, and $\mathbf{b}$ are model parameters.

In the process of training the ERNIE model, the input text token embeddings (Token Input) pass through N layers of the transformer (T-Encoder), after which, together with the named entity embeddings (Entity Input) they are processed by M layers of the aggregator (K-Encoder). At each layer i of the aggregator, the entity (e_1 and e_2) and text (w_1, …, w_n) embeddings pass through corresponding или related Multi-Head Attention block, and the corresponding updated entity ($\tilde{e}_1$ and $\tilde{e}_2$) and text ($\tilde{w}_1$, …, $\tilde{w}_n$) embeddings are fed into the knowledge injection block (Information Fusion), the output of which, according to the formulas (4), produces the embeddings of entities (Entity Output) and text (Token Output) taking into account the knowledge injection.

A similar method is at the heart of the KnowBert model [18], but the injection of external information occurs at the level of embeddings of entities, which are updated through the attention mechanism and by adding pretraining entity embeddings from the knowledge base, which subsequently also affects the embeddings of all tokens through the attention mechanism, according to the formula:

$$\mathbf{H}'_i = \mathrm{MLP}(\mathrm{MHA}(\mathbf{H}_i, \mathbf{S}'^e, \mathbf{S}'^e), \tag{5}$$

where $\mathbf{H}'_i$ is the embedding of the token i after knowledge injection, MLP is the multilayer perceptron, MHA is Multi-Head Attention, $\mathbf{H}_i$ is the embedding of the token i before knowledge injection, $\mathbf{S}'^e$ are updated embeddings of the identified named entities.

A conceptually similar architecture to ERNIE is proposed in [19], the main difference being the use of information about relations between entities, the prediction of which is represented by a separate task for pretraining. In the Weakly Supervised Knowledge-Pretrained Language Model [20], instead of masking entities in the pretraining phase, the model is trained to predict whether entities in the input data have been replaced by others of the same type within the Wikidata knowledge base. The architecture of the model in this case is consistent with BERT, but token masking is only performed at 5% rather than 15% to avoid masking too large a fragment of context, as entities can consist of multiple words. In [21], word combination masking is applied in addition to masking words and named entities, which improves the model's understanding of word combinability, and the injection is done in stages: at each stage, a BERT-like model is trained on only one type of masking. The use of multiple training modes, in which the model switches from word prediction to phrase prediction depending on which mode between the last two consecutive iterations had the largest reduction in model loss relative to the total reduction in loss over all iterations, is a major innovation [22].

The authors of the study [23] pretrain a BERT-based model, aiming to learn masked entity prediction from their descriptions, as well as to converge embeddings of synonymous entity descriptions and distance antonymous ones, for which a special loss function is used:

$$L = -\sum \log \frac{f(\mathbf{h}_\mathrm{ori}, \mathbf{h}_\mathrm{syn})}{f(\mathbf{h}_\mathrm{ori}, \mathbf{h}_\mathrm{syn}) + f(\mathbf{h}_\mathrm{ori}, \mathbf{h}_\mathrm{ant})}, \tag{6}$$

where $f(\mathbf{h}_i, \mathbf{h}_j) = \exp(\mathbf{h}_i \mathbf{h}_j)$, $\mathbf{h}_\mathrm{ori}$ is the embedding of the masked entity description, $\mathbf{h}_\mathrm{syn}$ is the embedding of the synonymous entity description, $\mathbf{h}_\mathrm{ant}$ is the embedding of the antonymous entity description.

In order to deal with specific tasks this model is used in pair with BERT as an additional source of knowledge in the form of embeddings of identified entities, and their injection is performed through concatenation of model outputs with BERT outputs with optional application of attention mechanism to take into account the importance of data on specific entities. The use of the attention mechanism is considered for both output embeddings from the last model layers and output embeddings across model layers with averaging applied, and the result of the attention mechanism is concatenated with the output of the BERT model instead of the output of the auxiliary model. Linguistic features also play an important role in [24], in which an additional task for self-supervised learning is to determine the semantic similarity of a pair of words, while the model is trained by alternating between the BERT self-supervised learning task and the auxiliary task. A similar idea is presented in [25], where the model is trained to classify words into groups with similar meaning based on WordNet[11] data.

[11] A lexical database of the English language developed at Princeton University. https://wordnet.princeton.edu/. Accessed December 01, 2024.

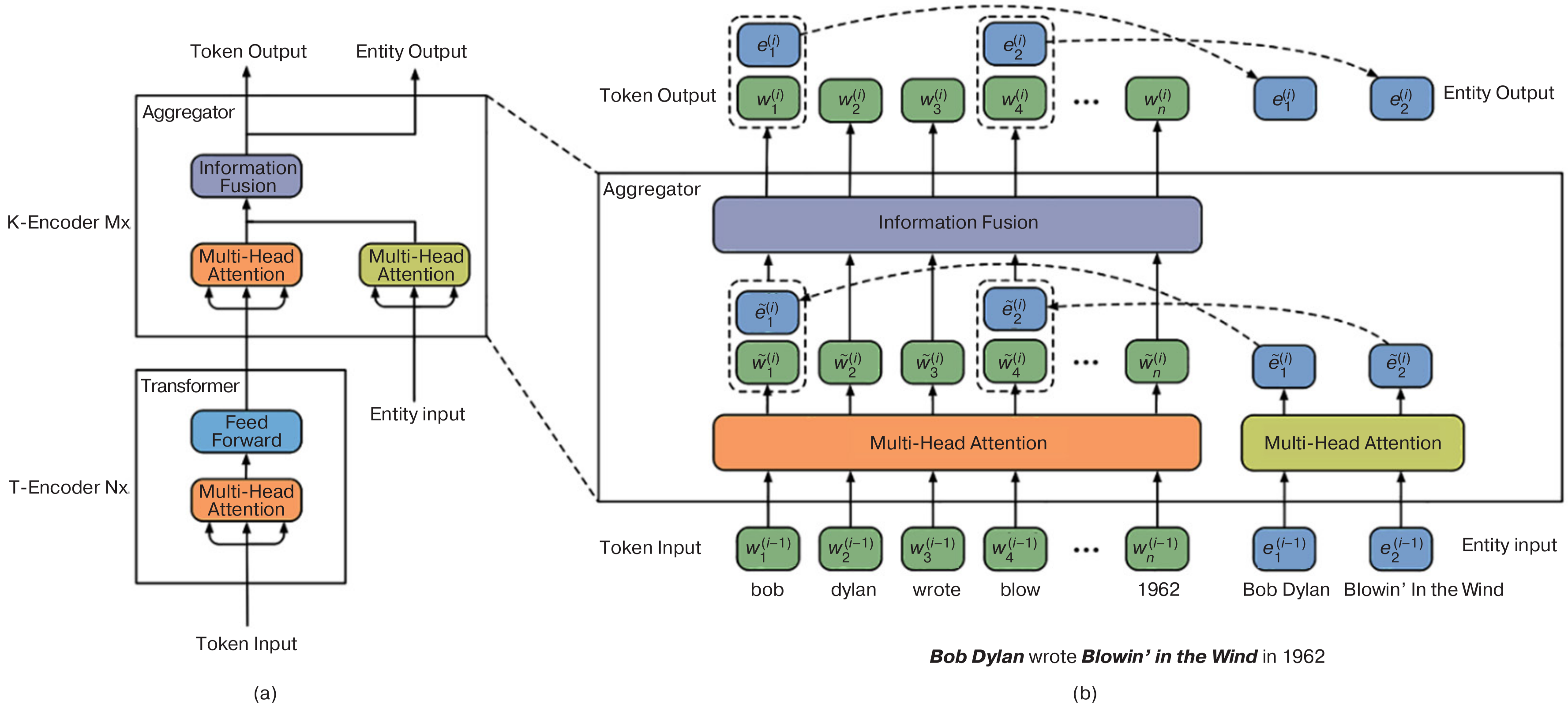


Fig. 3. ERNIE model training scheme [17]:
(a) model architecture; (b) aggregator architecture

A logical development of entity and relation masking approaches is the use of predicting structured knowledge units in the form of triplets as the pretraining stage task, which may enable learning more general principles and relationships similar to those contained in knowledge bases. Thus, in Knowledge Embedding and Pretrained Language Representation (KEPLER) [26], embeddings of triplet elements are treated as embeddings of their descriptions from a knowledge base, obtained using the same model that is used to generate embeddings of text tokens in the masked token prediction task. In this case, the scoring function of TrancE knowledge graph embedding model is applied to compute the loss in the triplet prediction task [27]:

$$d(\mathbf{h}, \mathbf{t}) = ||\mathbf{h} + \mathbf{r} - \mathbf{t}||, \tag{7}$$

where **h** is the embedding of the subject in the triplet, **r** is the embedding of the relation in the triplet, **t** is the embedding of the object in the triplet.

In [28], a set of triplets from a single subgraph is given as input to the model, and therefore the attention mechanism additionally utilizes the adjacency matrix to account for existing relationships, and training is performed in a triplet reconstruction format, which involves composing triplets from updated vertex embeddings and using the scoring function (7). The study [29] contains the idea of pretraining three functions based on an encoder model to predict each triplet element from the other two, which should facilitate the learning of possible combinations. In this setting, the answer score is the product of the similarity scores of the values of the three pretrained functions and the corresponding real triplet elements, where the subject represents the context of the question, the relation is the question itself, and the object is the specific answer choice. A pretraining function that will participate in finding answers to queries by identifying the most likely relationships with auxiliary data from the knowledge base is proposed in [30]. With the help of this function, the extracted auxiliary facts for each answer choice are compared by the degree of similarity with the facts for the question, and the more probable answer is considered to be the one for which this similarity of facts is higher on average.

At the same time, triplets from a relevant subgraph to a query can be directly fed to the input of the model within the framework of pretraining on a par with text tokens using special embeddings to indicate the token type, as shown in [31]. In this regard, when implementing the attention mechanism in the model, a mask matrix is used to restrict the interaction of unrelated vertices in the subgraph. In [32], it was proposed to improve the approach of the ERNIE model by modifying the representation of entities by taking into account their relationships in the corresponding subgraph, and using the attention mechanism to filter out potentially irrelevant context for a query.

Another way of using knowledge bases in the pretraining phase of the model can be to build new QA datasets on their basis, with which the system also improves its ability to find correct answers in a certain way. This approach is used in [33], and in [34] it is developed by conceptualization: specific facts are considered in a more general way, so that more situations can be covered and the ability to distinguish between similar variants can be improved. For example, using the ATOMIC framework, playing soccer can be represented as a tedious event.

The concept of Self-supervised Bidirectional Encoder Representation Learning of Commonsense (elBERto) model [35] is more emphasized on quantitative expansion of the number of self-supervised learning tasks. In order to improve the system's ability to process difficult queries, three more tasks were added to the BERT self-supervised learning tasks: the first one is aimed at distinguishing contexts with opposite meanings; the second one requires putting in order several jumbled sentences taken from the same paragraph; the third one extends the learning of contextual relationships through entity masking. According to the authors, this approach will also allow the system to better capture linguistic patterns and provide more universal applications.

Another concept of knowledge injection implies as an additional step the fine-tuning on the basis of existing datasets corresponding to the practical task. For example, the use of the SQuAD dataset [36] from the field of QA systems has gained some popularity in this regard. Its key features include a relatively large size (more than 100000 queries), while each query is accompanied by a corresponding context taken from Wikipedia. Thus, as a result of training on this dataset, the model better adapts to the problem formulation and format, and in addition processes a rather significant amount of data, thus increasing the amount of learned factual information.

As a relevant and characteristic example in this regard, we can mention the UnifiedQA model [37], the development of which was based on training the language model on 8 QA datasets of different types. It allows the existing benchmark formats to be adapted and provides an increase in the accuracy of the model on unseen questions in the training process, opening also new opportunities for its further fine-tuning. The feasibility of such an approach was also confirmed for the Unicorn model from [38], but in this case, the scope of the study was limited exclusively to CommonsenseQA datasets.

The methods considered above can be referred to the class of approaches without explicit involvement of knowledge bases, since the use of the corresponding systems does not imply the direct extraction of the context to the query exactly from knowledge bases, and the emphasis is created on the knowledge that was obtained in the process of training. In fact, the large language models developed in the last few years are essentially based on a similar idea: training on a large amount of qualitative data, taking into account different specificities and human preferences, can increase the versatility of the systems and the evaluation of the results obtained with their help, if this process is sufficiently scaled up.

The advantages of this class include, in a sense, greater versatility due to its independence from the use of knowledge bases. In addition, the significant reliance of the QA system architecture on pretrained and fine-tuned models allows us to simplify its development, subsequent use and adaptation to specific tasks.

At the same time, this class of approaches can be considered to a certain extent limited in its possibilities for further development. The point is that, in general, the increase in the efficiency of models here is associated with targeted and point improvement, expansion of the amount of learned information, while no fundamentally new mechanisms that improve the system's reasoning abilities are introduced. In addition, this direction does not practically solve the problem of the lack of interpretability of the received answers and their justification by the system, as well as the problem of knowledge obsolescence.

METHODS OF KNOWLEDGE INJECTION WITH THE KNOWLEDGE BASES

The basic unit of information in knowledge bases can be considered in terms of entity–relation–entity triplets, while more complex relationships can be conveyed by a set of triplets or paths. Knowledge base subgraphs used in addition to or instead of paths can be considered as a set of paths having common elements. As a result, in terms of information processing, a QA system may have 3 types of attributes in some combination:

1) features obtained by processing the query context by a language model;
2) features based on extracted paths;
3) features associated with subgraphs of the knowledge base.

Thus, one direction for research is how to effectively process these different types of features and how to combine the corresponding results.

One approach to injecting features into the system is based on the fact that triplets and their aggregates can often be translated quite easily into natural language sentences, and in this form, it is possible to feed them into the input of the language model as an auxiliary context. It should be noted, however, that in this form they can also serve as a justification for the resulting answer. An example of the implementation of such an approach is the Knowledge-Augmented language model PromptING (KAPING) [39]. In [40], its effectiveness is investigated in the context of using language models pretrained on auxiliary datasets. In the DEscriptive Knowledge for COmmonsense question answering (DEKCOR) model [41], in addition to the triplets extracted from ConceptNet, dictionary definitions of the corresponding entities are input, whereas in Knowledgeable External Attention for commonsense Reasoning (KEAR) [42] (Fig. 4) the context of the query is extended by including additional information from a number of QA datasets, which enables the use of more specific information.

In particular, under the KEAR architecture, to the concatenation of a question with one of the answer choices (Question & Candidate) relevant extracted auxiliary data (Knowledge Retrieval) from ConceptNet knowledge base, Wiktionary dictionary (Definition) and additional datasets (Training Data) are added to the model input. Embeddings of the query tokens ($E_{[\mathrm{CLS}]}$, E_0, …, E_N) to which a segment indicator (S_0) and the auxiliary context ($E_0^k, ..., E_{N_k}^k$) to which a segment indicator (S_1) is added are fed to the Transformer input. The answer probability (Score Prediction) is defined based on the final embedding of the auxiliary token $E_{[\mathrm{CLS}]}$ obtained by the attention mechanism (Self-Attention / External Attention).

The advantage of knowledge injection through language models is the possibility to rely heavily on the performance of pretrained models, while in some cases even avoiding the need to change their weights (e.g., the KAPING model). Also, the computational cost of acquiring additional features can be considered relatively small. At the same time, since most pretrained models can only efficiently utilize a fixed amount of information from the input data, there is a need to limit the amount of less relevant information.

In the simplest case, a restriction on the number of triplets (e.g., no more than 3 consecutive triplets) or a set of heuristics that take into account the peculiarities of a particular knowledge base can be used. In [43], in order to include only potentially relevant metadata from Wikidata in the model's input, the increase in the probability of a correct answer is estimated taking into account the corresponding Information Gain:

$$P(y \mid k_m) = 2^{\mathrm{pmi}(y,k_m)} P(y), \tag{8}$$

where y is the correct answer; k_m is a particular pattern containing m metadata; $\mathrm{pmi} = \log\left(\frac{P(y,k_m)}{P(y)P(k_m)}\right)$.

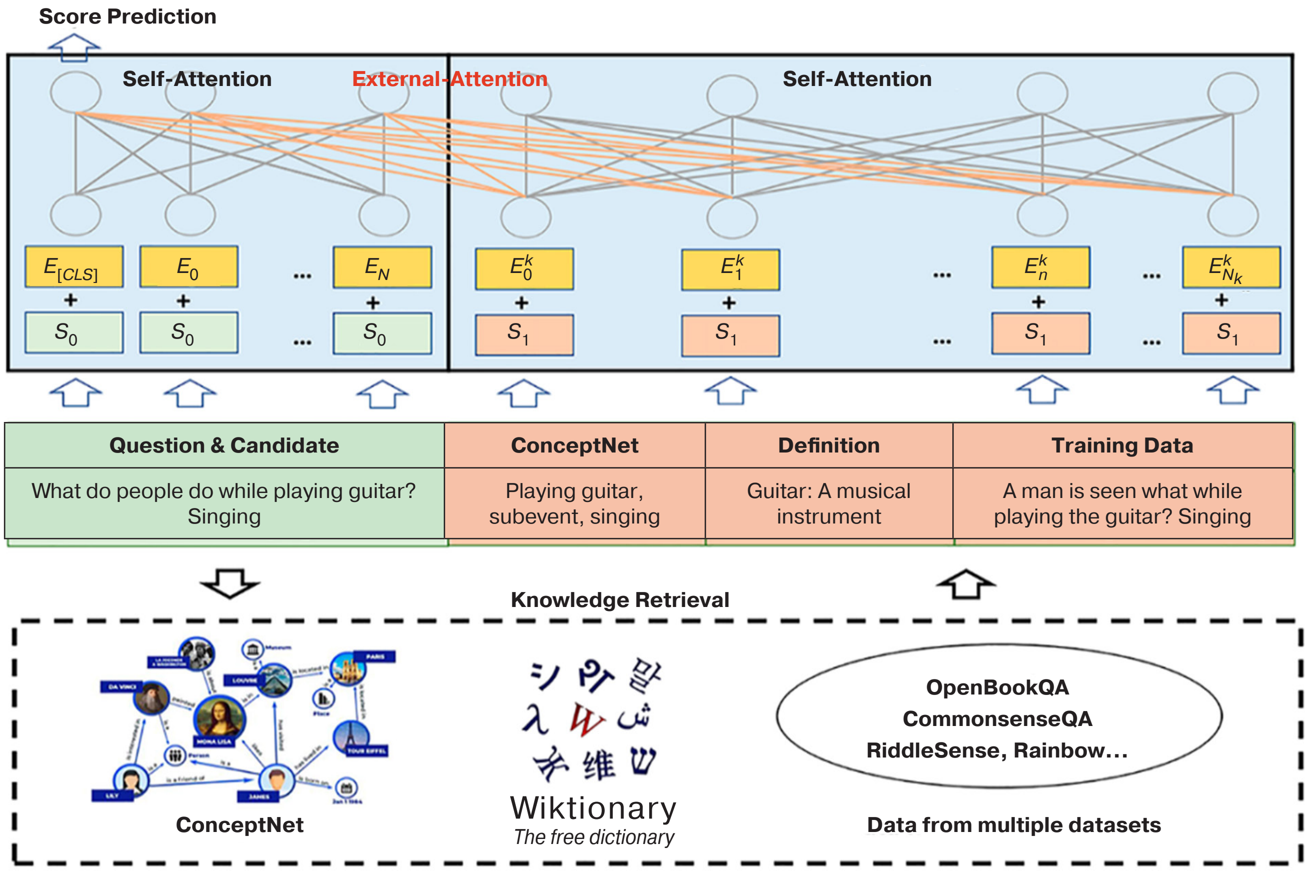


Fig. 4. KEAR model architecture [42]

The study [44] shows the feasibility of ranking the extracted additional information based on its importance estimation through an auxiliary pretrained model, while [45] shows the positive effect of using weighted summation of knowledge embeddings when injecting them, in addition to ranking, to emphasize more salient facts.

Also, in order to inject heterogeneous data, an attention mechanism can be applied to aggregate features based on their relevance to the task. In other words, the extracted knowledge that has more semantic relationship with the query, which is determined based on operations on embeddings, will be considered more relevant. Most often in practice, attention weights are derived based on operations on embeddings, which enable updating the relevant embeddings with respect to the specific task and its context. In particular, a similar approach is presented in works [46–48], and in the article [49] auxiliary knowledge is also filtered based on the frequency of occurrence of entities and relevant paths, while for knowledge injection a sigmoid function is additionally used to adjust how much they will affect the context update for the query. In addition to the attention mechanism for filtering out irrelevant data, the study [50] proposed to use graph-based approaches to determine the importance of individual nodes in the extracted subgraph: node closeness calculation, PageRank[12] and its modification, which enables only the most informative paths to be considered.

In addition, a disadvantage of knowledge injection through language models is the limited use of structured knowledge bases, which may reduce the potential efficiency of the final implementation. In order to preserve the effect of considering the relationships when translating triplets into text and to prevent information mixing in the K-BERT model [51], the positional encoding is included at the stage of generating embeddings, and in the subsequent computations, a specially introduced Visible Matrix is adopted, the elements of which determine what tokens a particular token should interact with in a given context.

In this regard, it is necessary to mention one of the main tools for processing structured knowledge—graph neural networks. This tool allows us to obtain and update embeddings of graph vertices using the concept of message passing:

$$\mathbf{h}_u = \phi(x_u, \bigoplus_{v \in N_u} \psi(x_u, x_v, e_{uv})), \qquad (9)$$

where $\mathbf{h}_u$ is the embedding of the vertex u; x_u and x_v are the features of the vertices u and v; e_{uv} is the feature of

[12] A ranking algorithm that evaluates the number and quality of links leading to web pages.

the edge between the vertices u and v; ϕ and ψ are the set differentiable functions; $\bigoplus_{v \in N_u}$ is a permutation invariant aggregation operator acting on the neighbors of the vertex u.

Due to this, in practice, the embedding of each entity can use different contextual data obtained from the knowledge base, taking into account in a certain way the information from its neighbors in the graph. An example of such a model used in the context of knowledge injection is the Graph Convolutional Network [52].

The features obtained by graph neural networks can also be subsequently injected into the system operation using an attention mechanism. Among the implementations of this kind is the model architecture [53] depicted in Fig. 5. Here, the embedding of each vertex of the auxiliary subgraph is adjusted for relevance with respect to the existing embedding of the query before it is directly used to obtain an answer:

$$\alpha_i = \frac{\mathbf{h}^{\mathrm{c}} \sigma(\mathbf{W}\mathbf{h}_i)}{\sum_{j \in N} \mathbf{h}^{\mathrm{c}} \sigma(\mathbf{W}\mathbf{h}_j)}, \tag{10}$$

where α_i is the relevance degree of the vertex i; $\mathbf{h}^{\mathrm{c}}$ is the embedding of the query context; $\mathbf{W}$ is the weight matrix; $\mathbf{h}_i$ is the embedding of the vertex i; N is the set of vertex indices neighboring the vertex i.

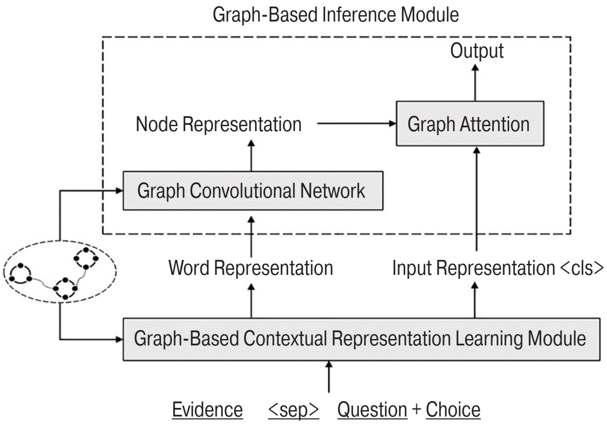

Graph-Based Inference Module
Output
Node Representation
Graph Attention
Graph Convolutional Network
Word Representation
Input Representation <cls>
Graph-Based Contextual Representation Learning Module
Evidence
<sep>
Question + Choice

Fig. 5. Model architecture from [53]

In general, the model is organized as follows: first, a subgraph with auxiliary information is extracted from an existing query consisting of a question and one of the answer choices. This information in the form of evidence is appended to the query and fed to the input of the language model (Graph-Based Contextual Representation Learning Module). The Word Representation token embeddings obtained at the output of the model are fed to the Graph-Based Inference Module, where they are used to initialize the corresponding nodes of the auxiliary graph, which are subsequently updated using the Graph Convolutional Network. The resulting Node Representation embeddings of the graph nodes are then aggregated using the Graph Attention mechanism with importance relative to the Input Representation embedding of the textual context, and the resulting graph embedding along with the textual embedding are directly used to predict the answer probability using a multilayer perceptron.

Similarly, the Multi-Hop Graph Relation Networks (MHGRN) model [54] is organized in a similar way, but its key difference is the consideration of the auxiliary subgraph as a set of paths connecting vertices, according to which the embedding of each vertex is updated based on the given length of paths from it. To aggregate information along the paths, special attention weights are introduced, which are defined as the conditional probability of a given sequence of triplets given the available context for a query, whereas to calculate the probability of a particular answer, the resulting embeddings of entities from the answer are aggregated using the attention mechanism and, together with the embedding of the query context, are processed by the multilayer perceptron. Thus, this approach also takes into account the importance of relations between entities. In the Joint reasoning with Language models and Knowledge graphs (JointLK) model [55], the least relevant nodes of the auxiliary subgraph are cut off and a new representation of the query context is additionally introduced, which takes into account the degree of its importance with respect to the subgraph and is the third component for obtaining the answer score along with the original context representation and the embedding of the subgraph. In the study [56], the message passing mechanism implements consistent updating of both entity and relation embeddings, which in this case are also used to estimate the answer probability. In this case, a modified adjacency matrix, whose elements are the corresponding attention weights, is used to formalize the relevance of relations between vertices under a given query context. In the Knowledge-Aware Graph Network (KagNet) module [57], the embeddings of the vertices of the auxiliary graph updated with the help of message passing mechanism are considered as elements of paths connecting entities from the question and one of the answer options. As a result, for each such pair of entities, a vector of structured features is generated as an average of the embeddings of the paths connecting them, and a vector of textual features obtained as the result of applying a multilayer perceptron to the concatenation of the embeddings of the query and each entity from the pair. To estimate the probability of a particular answer, averaging over all pairs of entities from the query and response is implemented. Furthermore, in addition, instead of averaging, the authors also propose to utilize the attention mechanism for feature aggregation.

One of the certain disadvantages of using graph neural networks is the increase in the number of parameters in the model and, consequently, in the resources for its training and use. In this regard, [58] proposes a simplified algorithm for obtaining triplet embeddings based on one-hot vectors indicating the type of entity in the graph and a certain relation within the ConceptNet database. To calculate the final answer probability, the model uses two scores: for textual and graph features. The former is based on the processing of the query embedding by the multilayer perceptron, while the latter is based on a weighted sum of path embeddings that takes into account their frequency of occurrence.

The process of knowledge injection may be somewhat more difficult when there are multiple sources of information and training on different types of tasks. In such conditions, it is necessary to solve the problems associated with the need to retrain the model weights and the displacement of learned facts by new ones, which can lead to unstable results. One possible solution is the use of adapters [59]—special modules oriented for a specific data source or task, which allows us not to change the weights of the main model and to train only a relatively small number of adapter weights, and thus avoid knowledge mixing. In practice, several different adapters are usually trained independently and then used together to solve a particular task. Thus, the model [60] employs two types of adapters: the first one is focused on learning general facts from knowledge bases, while the second one is focused on linguistic information. Within the architecture, each output from the transformer model layer is fed to the input of the corresponding adapter layer, resulting in the formation of certain auxiliary features on the last adapter layer, which can be used to predict the answer together with the outputs of the last transformer layer. In [61] (Fig. 6), a slightly different approach is implemented where the weights of the adapters pre-trained on data from ATOMIC, ConceptNet, WikiData, and WordNet knowledge bases are also not changed when training the model on a specific task, but instead knowledge injection is performed by the attention mechanism (formula), where the adapters form Value and Key, and the pretrained transformer forms Query:

At each model layer, input data passes through the transformer layer and enters the Zero-shot Fusion knowledge integration block both directly (circle 4) and after interaction with adapter models (circles 1, 2, and 3). In this block, embeddings interact within the attention mechanism (formula (1)): the output representation from the transformer are used as query, while the outputs from the adapters act as Value and Key. Subsequently, the result of the knowledge integration block is summed with the output from the Multi-Head Attention block of the transformer and normalized (Add & Norm). The goal

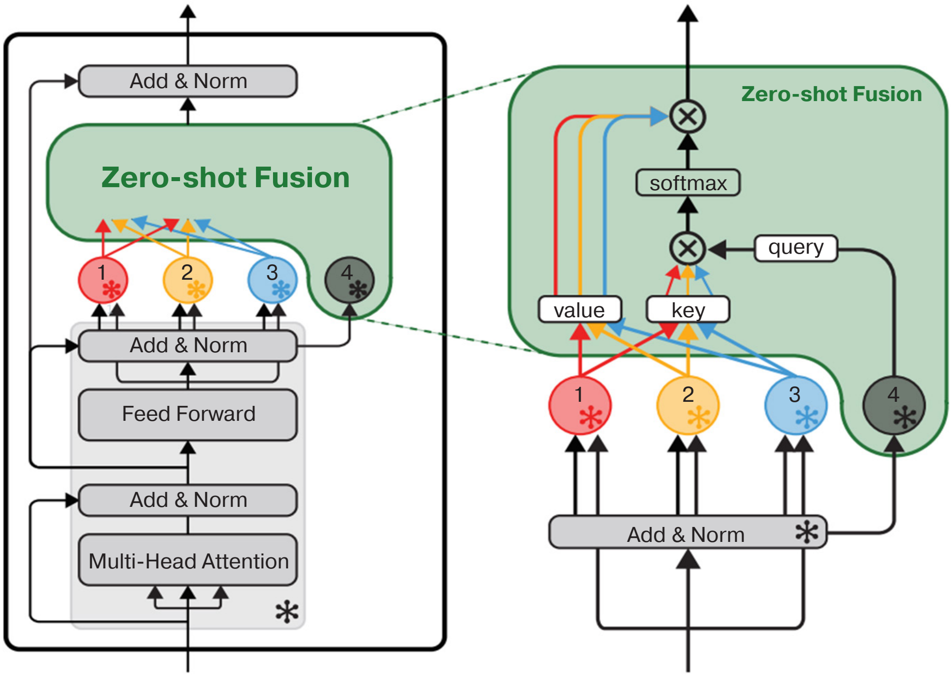


Fig. 6. Knowledge injection scheme using adapters from [61]

of model training under this architecture is to be able to address a more relevant adapter, which to some extent resembles the concept of mixture of experts [62].

Separately, in the context of knowledge injection, we can distinguish a group of approaches based on the use of so-called Interaction Tokens. The concept of interaction tokens is largely similar to the idea of using a special token [CLS] in language models, which can serve to classify a whole text fragment. Similarly, interaction tokens in the case of textual information or Interaction Nodes in the case of graphs can act as an intermediate container of necessary information for combining heterogeneous data. An example of the corresponding QA architecture can be seen in Fig. 7: within the Graph REASoning Enhanced Language Model (GreaseLM) [63] textual and structured information are processed independently, and their integration is realized by updating the vector representations of the interaction token and the interaction node by applying a bilayer perceptron to their concatenation:

$$[\mathbf{h}_{\text{int}};\mathbf{e}_{\text{int}}] = \text{MInt}([\tilde{\mathbf{h}}_{\text{int}};\tilde{\mathbf{e}}_{\text{int}}]) = \text{MLP}([\tilde{\mathbf{h}}_{\text{int}};\tilde{\mathbf{e}}_{\text{int}}]), \quad (11)$$

where $\tilde{\mathbf{h}}_{\text{int}}$ is the embedding of the interaction token before the knowledge integration, $\tilde{\mathbf{e}}_{\text{int}}$ is the embedding of the interaction vertex before the knowledge integration, MInt is the modality interaction layer.

In GreaseLM training, the concatenation of a question with one of the answer choices is processed using N layers of the model-encoder (Uni-modal Encoder) and together with the auxiliary graph (KG Retrieval) passes through M layers (GreaseLM Layer) of the knowledge integration block (Cross-modal Fuser). At each layer of this block, the embeddings of the text tokens (h_1, …, h_T, h_{int}) and graph vertices (e_{int}, e_1, …, e_J) are processed by the language model layer (LM Layer) and graph neural network layer (GNN Layer), respectively, and the integration process itself is carried out through the interaction (MInt, formula (11)) of the embeddings of the special tokens ($\tilde{h}_{\text{int}}$ and $\tilde{e}_{\text{int}}$). After the knowledge integration process is completed, the embeddings of the special tokens with the graph embedding (Pooling) obtained by means of the attention mechanism are used for Answer Selection by the perceptron (MLP).

Within the DRAGON[13] model [64], the GreaseLM architecture was considered in the context of self-supervised learning: after a knowledge integration layer, the obtained textual features are used to predict masked text tokens, while the graph features are used for the Link Prediction task, which involves establishing probability of a link between vertices in a graph using scoring functions similar to (7).

The Question Answer Graph Neural Network (QA-GNN) model [65] uses only an interaction node initialized by a embedding of the textual context from the query, based on similarity with which, determined using a pretrained model, the relevance of other nodes is estimated. These evaluations, together with features representing the types of vertices and relations in the form of one-hot encoding, are used to compute attention weights, which are used to implement the message passing between vertices and the corresponding update of their embeddings. The answer selection process is also essentially formulated similarly to the GreaseLM model. In PipeNet [66], compared to QA-GNN, the computation of the relevance of the vertices of the auxiliary graph to the query context is, in a sense, replaced by an algorithm for cutting off irrelevant vertices, based on determining the shortest distance between entities within the language dependency graph corresponding to the query:

$$D(c_{\text{q}}) = -\frac{\sum_{i=1}^{|V_{\text{a}}|} Dist(c_{\text{q}}, c_{\text{a}})}{|V_{\text{a}}|}, \quad (12)$$

where $D(c_{\text{q}})$ is the relevance of the entity c_{q} from the query, $Dist(c_{\text{q}}, c_{\text{a}})$ is the shortest distance between the entity c_{q} from the query and the entity c_{a} from the corresponding answer choice, V_{a} is the set of entities from the answer choice for the query.

The rest of the QA-GNN architecture is essentially the same, except for the use of vertex relevance scores in the calculation of attention weight.

To summarize the methods of knowledge injection with graph models, it can be stated that they are characterized by the greatest variety of ideas used, which demonstrate a wide range of possibilities for taking into account the features of structured knowledge and their inclusion in the work of QA systems. A positive aspect can also be considered the possibility to increase the interpretability of the model due to the formation of fact chains with the help of knowledge bases, which can be updated separately in a timely manner depending on the current events. At the same time, the full-fledged integration of graph features leads to a significant complication of model architectures and, depending on the implementation, may require certain additional computational resources, as a result of which the benefit of knowledge injection becomes more ambiguous.

[13] DRAGON— Deep Bidirectional Language-Knowledge Graph Pretraining.

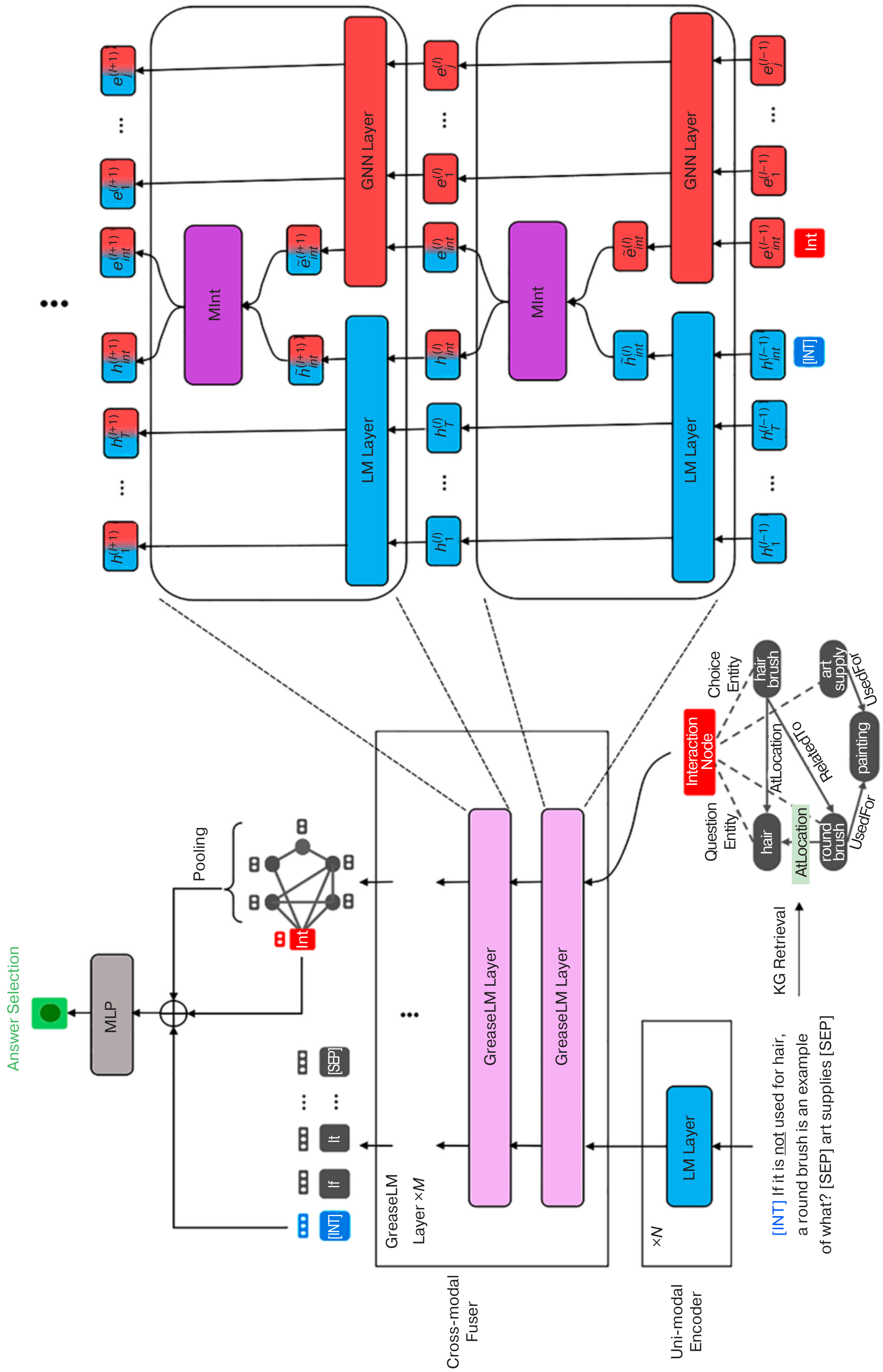


Fig. 7. GreaseLM model architecture [63]

COMPARATIVE ANALYSIS

The described approaches also differ from the point of view of setting up the experimental part. Thus, first of all, different benchmarks could be used to test the efficiency of implementations. As a result, it was decided to perform the comparative analysis with respect to the CommonsenseQA dataset that appeared most frequently in the considered works [67].

CommonsenseQA consists of 12102 questions offering five answer choices, one of which is correct. The choice in favor of this dataset can be justified by its higher complexity in terms of the relatively poor results of QA systems on it compared to its counterparts. In this case, the higher complexity is due to the focus of the questions on social and psychological aspects and the need to establish causal relationships, as well as the lack of any additional context for the questions. While this complicates the effective implementation of pretrained language models due to the smaller number of inputs, such a formulation of the problem favors the formation of such a context through external knowledge bases.

Table 2 summarizes the results of the models without Ensemble on the CommonsenseQA dataset test sample. For practical comparison of implementations in the context of this dataset, accuracy (the percentage of questions that were answered correctly) is used as a metric. It should also be noted that one of the most frequently used language encoder models, RoBERTa [68], was chosen as the baseline benchmark.

The results show that any of the considered approaches can increase the accuracy of the QA system with respect to the base solution using a pretrained language model, thus confirming the promising avenue of this line of

Table 2. Comparison of the effectiveness of knowledge injection methods

Model	Injection method	Accuracy on CommonsenseQA test set, %
RoBERTa [68] (2019)	–	68.7
Model from [15] (2020)	Self-supervised learning	75.6
Model from [23] (2022)	Self-supervised learning	78.5
UnifiedQA [37] (2020)	Fine-tuning	79.1
Model from [44] (2023)	Text embeddings and attention mechanism	75.0
Model from [47] (2020)	Text embeddings and attention mechanism	80.3
DEKCOR [41] (2021)	Text embeddings and attention mechanism	80.7
KEAR [42] (2022)	Text embeddings and attention mechanism	86.1
JointLK [55] (2022)	Graph embeddings and attention mechanism	74.4
Modelfrom [53] (2020)	Graph embeddings and attention mechanism	75.3
MHGRN [54] (2020)	Graph embeddings and attention mechanism	75.4
QA-GNN [65] (2021)	Interaction tokens	73.4
GreaseLM [63] (2022)	Interaction tokens	74.2
DRAGON [64] (2022)	Interaction tokens	76.0

research. At the same time, the models using graph-based embeddings demonstrate noticeably lower accuracy, while the best result on the CommonsenseQA dataset is obtained using the KEAR model with knowledge base information injection via text-based embeddings.

However, from a practical point of view, the existence of other important factors should be taken into account when comparing approaches. For example, models based on self-supervised learning and fine-tuning, despite their lower accuracy, require less additional computations to obtain an answer to a query. At the same time, the very process of pretraining such models implies a rather significant expenditure of computational resources. In addition, not all implementations use the same language models, which in itself may result in differences in the final accuracy. The amount of time the model needs to obtain an answer it can also be considered as a relevant factor.

If proceeding solely from the results on the CommonsenseQA benchmark, it can be stated that the use of more architecturally complex models in general does not have a significant enough effect to compete with more established approaches that focus solely on the use of language models. Nevertheless, it continues to be worthwhile to continue the comparative analysis using other benchmarks as a means of better assessing the real state of art.

CONCLUSIONS

The presented review forms a basis to argue for the effectiveness of knowledge injection techniques in the field of QA system design. Already existing solutions experimentally confirm the possibility of simultaneously achieving several main goals of knowledge injection in this context.

However, there is still considerable room for further improvement in multiple aspects of the process. Firstly, currently relatively basic and well-established in the field of natural language processing methods for extracting data from knowledge bases for a query prevail. Only a few works propose ways to improve this process, such as paraphrasing additional knowledge from the database to simplify its processing by the system. In this context, given the potential importance of extracting relevant information in terms of further implementation, specific approaches can be considered along with their impact on the result.

Secondly, it is of interest to analyze the potential impact of choosing a particular graph model for processing structured information, since in existing works the main emphasis is shifted to comparison according to the criterion of the used language model. At the same time, over the last few years, many new promising models of knowledge graphs embeddings and graph neural networks have appeared, whose capabilities in the framework of practical tasks of this kind have yet to be established, but can significantly affect the results of the system as a whole.

Thirdly, there is currently a lack of systematic studies comparing methods for combining data from different modalities in the context of QA system design. This issue can also be considered relevant due to the possibility of generalizing to a wider range of tasks.

Finally, within the current vector of development of the field of QA system design leading to the prevalence of universal generative language models such as ChatGPT in applications, it makes sense to emphasize the study of the peculiarities of knowledge injection methods in this type of model.

REFERENCES

1. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2019;1:4171–4186. https://doi.org/10.18653/v1/N19-1423
2. Petroni F., Rocktäschel T., Lewis P., et al. Language Models as Knowledge Bases? *Processing* (*EMNLP-IJCNLP*). 2019. P. 2463–2473. https://doi.org/10.18653/v1/D19-1250
3. Sap M., Le Bras R., Allaway E., et al. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019;33(1):3027–3035. https://doi.org/10.1609/aaai.v33i01.33013027
4. Niven T., Kao H.-Y. Probing Neural Network Comprehension of Natural Language Arguments. *arXiv preprint arXiv:1907.07355*. 2019. https://doi.org/10.48550/arXiv.1907.07355
5. McCoy R. T., Pavlick E., Linzen T. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019. P. 3428–3448. http://doi.org/10.18653/v1/P19-1334
6. Li J., Chen J., Ren R., et al. The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models. *arXiv preprint arXiv:2401.03205*. 2024. https://doi.org/10.48550/arXiv.2401.03205
7. Wei J., Wang X., Schuurmans D., et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: *36th Conference on Neural Information Processing Systems*. 2022;35:24824–24837. https://doi.org/10.48550/arXiv.2201.11903
8. Lewis P., Perez E. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*. 2020;33:9459–9474. https://doi.org/10.48550/arXiv.2005.11401
9. Ye Zhi-Xiu, Chen Q., Wang W., Ling Zhen-Hua. Align, Mask and Select: A Simple Method for Incorporating Commonsense Knowledge into Language Representation Models. *arXiv preprint arXiv:1908.06725v5*. 2020. https://doi.org/10.48550/arXiv.1908.06725
10. Vaswani A., Shazeer N., Parmar N., et al. Attention Is All You Need. *Advances in Neural Information Processing Systems 30*. 2018. https://doi.org/10.48550/arXiv.1706.03762
11. Liu J., Shen D., Zhang Y., et al. What Makes Good In-Context Examples for GPT-3? *arXiv preprint arXiv:2101.06804*. 2021. https://doi.org/10.48550/arXiv.2101.06804
12. Gao T., Fisch A., Chen D. Making Pre-trained Language Models Better Few-shot Learners. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (*Volume 1: Long Papers*). 2021. P. 3816–3830. http://doi.org/10.18653/v1/2021.acl-long.295
13. Shwartz V., West P., Le Bras R., et al. Unsupervised Commonsense Question Answering with Self-Talk. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*). 2020. P. 4615–4629. http://doi.org/10.18653/v1/2020.emnlp-main.373
14. Wang J., Zhao H. ArT: All-round Thinker for Unsupervised Commonsense Question-Answering. In: *Proceedings of the 29th International Conference on Computational Linguistics*. 2022. P. 1490–1501. https://doi.org/10.48550/arXiv.2112.13428
15. Wang P., Peng N., Ilievski F., et al. Connecting the Dots: A Knowledgeable Path Generator for Commonsense Question Answering. *arXiv preprint arXiv:2005.00691*. 2020. https://doi.org/10.48550/arXiv.2005.00691
16. Raffel C., Shazeer N., Roberts A., et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*. 2020;21(140):1–67. https://doi.org/10.48550/arXiv.1910.10683
17. Zhang Z., Han X., Liu Z., et al. ERNIE: Enhanced Language Representation with Informative Entities. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019. P. 1441–1451. https://doi.org/10.18653/v1/P19-1139
18. Peters M.E., Neumann M., Logan IV R.L., et al. Knowledge enhanced contextual word representations. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*). 2019. P. 43–54. https://doi.org/10.18653/V1/D19-1005
19. He L., Zheng S., Yang T., Zhang F. KLMo: Knowledge Graph Enhanced Pretrained Language Model with Fine-Grained Relationships. In: *Findings of the Association for Computational Linguistics: EMNLP*. 2021. P. 4536–4542. https://doi.org/10.18653/v1/2021.findings-emnlp.384
20. Xiong W., Du J., Wang W.Y., Stoyanov V. Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model. *arXiv preprint arXiv:1912.09637*. 2019. https://doi.org/10.48550/arXiv.1912.09637
21. Sun Y., Wang S., Li Y., et al. ERNIE: Enhanced Representation through Knowledge Integration. *arXiv preprint arXiv:1904.09223*. 2019. https://doi.org/10.48550/arXiv.1904.09223
22. Zhang D., Yuan Z., Liu Y., et al. E-BERT: A Phrase and Product Knowledge Enhanced Language Model for E-commerce. *arXiv preprint arXiv:2009.02835*. 2020. https://doi.org/10.48550/arXiv.2009.02835
23. Chen Q., Li F.-L., Xu G., et al. DictBERT: Dictionary Description Knowledge Enhanced Language Model Pre-training via Contrastive Learning. *arXiv preprint arXiv:2208.00635*. 2022. https://doi.org/10.48550/arXiv.2208.00635

24. Lauscher A., Vulić I., Ponti E.M., et al. Informing Unsupervised Pretraining with External Linguistic Knowledge. *arXiv preprint arXiv:1909.02339v1*. 2019. https://doi.org/10.48550/arXiv.1909.02339
25. Levine Y., Lenz B., Dagan O., et al. SenseBERT: Driving Some Sense into BERT. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020. P. 4656–4667. https://doi.org/10.18653/v1/2020.acl-main.423
26. Wang X., Gao T., Zhu Z., et al. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Trans. Assoc. Comput. Linguis*. 2021;9:176–194. https://doi.org/10.1162/tacl_a_00360
27. Bordes A., Usunier N., Garcia-Durán A., et al. Translating Embeddings for Modeling Multi-relational Data. *Advances in Neural Information Processing Systems*. 2013. P. 2787–2795.
28. He B., Zhou D., Xiao J., et al. BERT-MK: Integrating Graph Contextualized Knowledge into Pre-trained Language Models. *Findings of the Association for Computational Linguistics: EMNLP*. 2020. P. 2281–2290. https://doi.org/10.18653/v1/2020.findings-emnlp.207
29. Banerjee P., Baral C. Self-Supervised Knowledge Triplet Learning for Zero-shot Question Answering. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*). 2020. P. 151–162. https://doi.org/10.18653/v1/2020.emnlp-main.11
30. Zhong W., Tang D., Duan N., et al. Improving Question Answering by Commonsense-Based Pre-training. In: Tang J., Kan M.Y., Zhao D., Li S., Zan H. (Eds.). *Natural Language Processing and Chinese Computing. NLPCC 2019. Lecture Notes in Computer Science*. Springer; 2019. V. 11838. P. 16–28. https://doi.org/10.1007/978-3-030-32233-5_2
31. Sun T., Shao Y., Qiu X., et al. CoLAKE: Contextualized Language and Knowledge Embedding. *arXiv preprint arXiv:2010.00309v1*. 2020. https://doi.org/10.48550/arXiv.2010.00309
32. Su Y., Han X., Zhang Z., et al. CokeBERT: Contextual knowledge selection and embedding towards enhanced pre-trained language models. *AI Open*. 2021;2:127–134. https://doi.org/10.1016/j.aiopen.2021.06.004
33. Ma K., Ilievski F., Francis J., et al. Knowledge-driven Data Construction for Zero-shot Evaluation in Commonsense Question Answering. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021;35(15):13507–13515. https://doi.org/10.1609/aaai.v35i15.17593
34. Wang W., Fang T., Ding W., et al. CAR: Conceptualization-Augmented Reasoner for Zero-Shot Commonsense Question Answering. *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023. P. 13520–13545. https://doi.org/10.18653/v1/2023.findings-emnlp.902
35. Zhan X., Li Y., Dong X., et al. elBERto: Self-supervised Commonsense Learning for Question Answering. *arXiv preprint arXiv:2203.09424v1*. 2022. https://doi.org/10.48550/arXiv.2203.09424
36. Rajpurkar P., Jia R., Liang P. Know What You Don’t Know: Unanswerable Questions for SQuAD. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2018;2(Short Papers):784–789. https://doi.org/10.18653/v1/P18-2124
37. Khashabi D., Min S., Khot T., et al. UnifiedQA: Crossing Format Boundaries with a Single QA System. In: *Findings of the Association for Computational Linguistics*. 2020. P. 1896–1907. https://doi.org/10.18653/v1/2020.findings-emnlp.171
38. Lourie N., Le Bras R., Bhagavatula C., Choi Y. UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark. *arXiv preprint arXiv:2103.13009v1*. 2021. https://doi.org/10.48550/arXiv.2103.13009
39. Baek J., Aji A.F., Saffari A. Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering. In: *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations* (*NLRSE*). 2023. P. 78–106. https://doi.org/10.18653/v1/2023.nlrse-1.7
40. Pan X., Sun K., Yu D., et al. Improving Question Answering with External Knowledge. In: *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. 2019. P. 27–37. https://doi.org/10.18653/v1/D19-5804
41. Xu Y., Zhu C., Xu R., et al. Fusing Context Into Knowledge Graph for Commonsense Question Answering. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021. P. 1201–1207. https://doi.org/10.18653/v1/2021.findings-acl.102
42. Xu Y., Zhu C., Wang S., et al. Human Parity on CommonsenseQA: Augmenting Self-Attention with External Attention. In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence* (*IJCAI*). 2022. P. 2762–2768. https://doi.org/10.24963/ijcai.2022/383
43. Arora S., Wu S., Liu E., Ré C. Metadata Shaping: A Simple Approach for Knowledge-Enhanced Language Models. In: *Findings of the Association for Computational Linguistics: ACL 2022*. 2022. P. 1733–1745. https://doi.org/10.18653/v1/2022.findings-acl.137
44. Li S., Gao Y., Jiang H., et al. Graph Reasoning for Question Answering with Triplet Retrieval. In: *Findings of the Association for Computational Linguistics: ACL 2023*. 2023. P. 3366–3375. https://doi.org/10.18653/v1/2023.findings-acl.208
45. Mitra A., Banerjee P., Pal K.K., et al. How Additional Knowledge can Improve Natural Language Commonsense Question Answering? *arXiv preprint arXiv:1909.08855v3*. 2020. https://doi.org/10.48550/arXiv.1909.08855
46. Chen Q., Zhu X., Ling Z.-H., et al. Neural Natural Language Inference Models Enhanced with External Knowledge. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* 2018;1(Long Papers):2406–2417. https://doi.org/10.18653/v1/P18-1224

47. Chen Q., Ji F., Chen H., Zhang Y. Improving Commonsense Question Answering by Graph-based Iterative Retrieval over Multiple Knowledge Sources. In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020. P. 2583–2594. https://doi.org/10.18653/v1/2020.coling-main.232
48. Ma K., Francis J., Lu Q., et al. Towards Generalizable Neuro-Symbolic Systems for Commonsense Question Answering. In: *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*. 2019. P. 22–32. https://doi.org/10.18653/v1/D19-6003
49. Bauer L., Wang Y., Bansal M. Commonsense for Generative Multi-Hop Question Answering Tasks. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018. P. 4220–4230. https://doi.org/10.18653/v1/D18-1454
50. Paul D., Frank A. Ranking and Selecting Multi-Hop Knowledge Paths to Better Predict Human Needs. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 2019;1(Long and Short Papers):3671–3681. https://doi.org/10.18653/v1/N19-1368
51. Liu W., Zhou P., Zhao Z., et al. K-BERT: Enabling Language Representation with Knowledge Graph. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020;34(03):2901–2908. https://doi.org/10.1609/aaai.v34i03.5681
52. Kipf T.N., Welling M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907*. 2017. https://doi.org/10.48550/arXiv.1609.02907
53. Lv S., Guo D., Xu J., et al. Graph-Based Reasoning over Heterogeneous External Knowledge for Commonsense Question Answering. In: *Proceedings of the AAAI Conference on Artificial Intelligence.* 2020;34(05):8449–8456. https://doi.org/10.1609/aaai.v34i05.6364
54. Feng Y., Chen Y., Lin B.Y., et al. Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*). 2020. P. 1295–1309. https://doi.org/10.18653/v1/2020.emnlp-main.99
55. Sun Y., Shi Q., Qi L., Zhang Y. JointLK: Joint Reasoning with Language Models and Knowledge Graphs for Commonsense Question Answering. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2022. P. 5049–5060. https://doi.org/10.18653/v1/2022.naacl-main.372
56. Yan J., Raman M., Chan A., et al. Learning Contextualized Knowledge Structures for Commonsense Reasoning. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP*. 2021. P. 4038–4051. https://doi.org/10.18653/v1/2021.findings-acl.354
57. Lin B.Y., Chen X., Chen J., Ren X. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*). 2019. P. 2829–2839. https://doi.org/10.18653/v1/D19-1282
58. Jiang J., Zhou K., Zhao W.X., Wen J.-R. Great Truths are Always Simple: A Rather Simple Knowledge Encoder for Enhancing the Commonsense Reasoning Capacity of Pre-Trained Models. In: *North American Chapter of the Association for Computational Linguistics-Findings*. 2022. https://doi.org/10.48550/arXiv.2205.01841
59. Houlsby N., Giurgiu A., Jastrzebski S., et al. Parameter-Efficient Transfer Learning for NLP. In: *Proceedings of Machine Learning Research*. 2019;97:2790–2799. https://doi.org/10.48550/arXiv.1902.00751
60. Wang R., Tang D., Duan N., et al. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP.* 2021. P. 1405–1418. https://doi.org/10.18653/v1/2021.findings-acl.121
61. Kim Y.J., Kwak B., Kim Y., et al. Modularized Transfer Learning with Multiple Knowledge Graphs for Zero-shot Commonsense Reasoning. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2022. P. 2244–2257. https://doi.org/10.18653/v1/2022.naacl-main.163
62. Jacobs R., Jordan M., Nowlan S., Hinton G. Adaptive Mixtures of Local Experts. *Neural Computation*. 1991;3(1):79–87. https://doi.org/10.1162/neco.1991.3.1.79
63. Zhang X., Bosselut A., Yasunaga M., et al. GreaseLM: Graph REASoning Enhanced Language Models for Question Answering. In: *The International Conference on Learning Representations* (*ICLR*). 2022. https://doi.org/10.48550/arXiv.2201.08860
64. Yasunaga M., Bosselut A., Ren H., et al. Deep Bidirectional Language-Knowledge Graph Pretraining. In: *36th Conference on Neural Information Processing Systems* (*NeurIPS*). 2022. https://doi.org/10.48550/arXiv.2210.09338
65. Yasunaga M., Ren H., Bosselut A., et al. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 2021. P. 535–546. https://doi.org/10.18653/v1/2021.naacl-main.45

66. Su Y., Zhang J., Song Y., Zhang T. PipeNet: Question Answering with Semantic Pruning over Knowledge Graphs. *arXiv preprint arXiv:2401.17536v2*. 2024. https://doi.org/10.48550/arXiv.2401.17536

67. Talmor A., Herzig J., Lourie N., Berant J. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In*: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2019;1(Long and Short Papers):4149–4158. https://doi.org/10.18653/v1/N19-1421

68. Liu Y., Ott M., Goyal N., et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*. 2019. https://doi.org/10.48550/arXiv.1907.11692

69. Robertson S.E., Walker S. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In: *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. 1994. P. 232–241. https://doi.org/10.1007/978-1-4471-2099-5_24

About the Author

Daniil V. Radyush, Postgraduate Student, Faculty of Software Engineering and Computer Systems, ITMO University (49-A, Kronverkskii pr., Saint Petersburg, 197101 Russia). E-mail: daniil.radyush@gmail.com. Scopus Author ID 58234958500, https://orcid.org/0000-0001-8823-0609

Об авторе

Радюш Даниил Валентинович, аспирант, факультет программной инженерии и компьютерной техники, ФГАОУ ВО «Национальный исследовательский университет ИТМО» (197101, Россия, Санкт-Петербург, Кронверкский пр., д. 49, лит. А). E-mail: daniil.radyush@gmail.com. Scopus Author ID 58234958500, https://orcid.org/0000-0001-8823-0609

Translated from Russian into English by L. Bychkova
Edited for English language and spelling by Thomas A. Beavitt

Information systems. Computer sciences. Issues of information security

Информационные системы. Информатика. Проблемы информационной безопасности

UDC 004.2
https://doi.org/10.32362/2500-316X-2025-13-3-44-53
EDN QWXGNC

RESEARCH ARTICLE

Method for designing specialized computing systems on the basis of hardware and software cooptimization

Ilya E. Tarasov @,
Peter N. Sovietov,
Daniil V. Lulyava,
Nikita A. Duksin

MIREA – Russian Technological University, Moscow, 119454 Russia
@ Corresponding author, e-mail: tarasov_i@mirea.ru

• **Submitted:** 03.11.2024 • **Revised:** 13.02.2025 • **Accepted:** 20.03.2025

Abstract

Objectives. Pipelining is an effective method for increasing the clock frequency of digital circuits. At the same time, balancing the pipeline stages during circuit synthesis at the register transfer level does not yet guarantee a balanced topological implementation of such a pipeline in terms of signal propagation delays according to the selected technological basis. This is due to the specifics of the algorithms for placing and routing components of digital devices, which are not capable of optimizing solutions in a strict mathematical sense in an acceptable time. In practice, approaches for developing digital devices combine manual control of topological constraints that set general rules for placing components with automatic optimization for localized fragments of the circuit are used to obtain results close to optimal. Pipeline circuits are based on a simple connection diagram of individual stages to demonstrate the effect of using topological design constraints on their example. On the basis of pipeline structures, a number of algorithms can be implemented to effectively complement programmable processor devices and provide hardware acceleration of some tasks. The present work develops methodological recommendations for managing topological design constraints in the implementation of pipeline computing structures based on programmable logic devices (PLD) with field-programmable gate array (FPGA) architecture.

Methods. The work is based on accepted methods for designing and modeling digital systems.

Results. Based on the analysis, modifications to a 32-bit CORDIC transcendental function computation pipeline were developed. By adding design constraints on the placement of register groups corresponding to the pipeline stages a significant increase in the clock frequency can be achieved as compared to automatic placement to reduce the running time of the tracing algorithms. The resulting effect is systematically reproduced in several implemented versions of the pipeline.

Conclusions. The presented recommendations can be used to control the clock frequency and number of stages of pipeline computing structures while simultaneously reducing the time of one iteration and routing of a module based on PLD with FPGA architecture.

Keywords: PLD, pipeline, constraints, CORDIC

© I.E. Tarasov, P.N. Sovietov, D.V. Lulyava, N.A. Duksin, 2025

For citation: Tarasov I.E., Sovietov P.N., Lulyava D.V., Duksin N.A. Method for designing specialized computing systems on the basis of hardware and software cooptimization. *Russian Technological Journal.* 2025;13(3):44–53. https://doi.org/10.32362/2500-316X-2025-13-3-44-53, https://www.elibrary.ru/QWXGNC

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Управление топологическими ограничениями при реализации конвейерных вычислительных структур на базе программируемых логических интегральных схем

И.Е. Тарасов [@]**,**
П.Н. Советов,
Д.В. Люлява,
Н.А. Дуксин

МИРЭА – Российский технологический университет, Москва, 119454 Россия
[@] *Автор для переписки, e-mail: tarasov_i@mirea.ru*

• **Поступила:** 03.11.2024 • **Доработана:** 13.02.2025 • **Принята к опубликованию:** 20.03.2025

Резюме

Цели. Конвейеризация является эффективным приемом повышения тактовой частоты цифровых схем. При этом балансировка стадий конвейера при синтезе схемы на уровне регистровых передач еще не гарантирует сбалансированную по задержкам распространения сигнала топологическую реализацию такого конвейера в выбранном технологическом базисе. Это обусловлено спецификой алгоритмов размещения и трассировки компонентов цифровых устройств, которые не позволяют получать оптимальные решения в строгом математическом смысле за приемлемое время. В практике разработки цифровых устройств применяются подходы, основанные на комбинации ручного управления топологическими ограничениями, задающими общие правила размещения компонентов, и автоматической оптимизации для локализованных фрагментов схемы, которая в этом случае позволяет получать результаты, близкие к оптимальным. Конвейерные структуры имеют простую схему соединений отдельных стадий, что позволяет продемонстрировать на их примере эффект от применения топологических проектных ограничений. В то же время, на базе конвейерных структур возможна реализация ряда алгоритмов, эффективно дополняющих программируемые процессорные устройства и обеспечивающие аппаратное ускорение некоторых задач. Цель работы – разработка методических рекомендаций по управлению топологическими проектными ограничениями при реализации конвейерных вычислительных структур на базе программируемых логических интегральных схем (ПЛИС) с архитектурой field-programmable gate array (FPGA).

Методы. Использованы методы проектирования и моделирования цифровых систем.

Результаты. На основе проведенного анализа разработаны модификации конвейерного вычислителя 32-разрядного преобразования CORDIC для вычисления трансцендентных функций. Установлено, что добавление проектных ограничений по размещению групп регистров, соответствующих стадиям конвейера, позволяет существенно повысить тактовую частоту по сравнению с автоматическим размещением и уменьшить время работы алгоритмов трассировки. Полученный эффект систематически воспроизводится в нескольких реализованных вариантах конвейера.

Выводы. Рассмотренные рекомендации позволяют управлять тактовой частотой и количеством стадий конвейерных вычислительных структур при одновременном уменьшении времени одной итерации размещения и трассировки модуля на базе ПЛИС.

Ключевые слова: ПЛИС, конвейер, проектные ограничения, CORDIC

Для цитирования: Тарасов И.Е., Советов П.Н., Люлява Д.В., Дуксин Н.А. Управление топологическими ограничениями при реализации конвейерных вычислительных структур на базе программируемых логических интегральных схем. *Russian Technological Journal*. 2025;13(3):44–53. https://doi.org/10.32362/2500-316X-2025-13-3-44-53, https://www.elibrary.ru/QWXGNC

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

High performance computing systems are designed as a combination of general purpose and specialized subsystems. At the architectural design stage, it is necessary to identify tasks to be solved by specialized subsystems that complement the work of general purpose processors. Such identified tasks should be both in high demand and either too inefficient to be solved by the central processing unit or represent an unacceptable load. In digital electronic devices and systems, algorithms for digital signal processing [1], calculation of hash functions in information protection subsystems [2], acceleration of artificial neural networks [3], etc., are often implemented based on specialized computing devices. In the present paper, approaches to the design of a pipeline calculator are considered on a number of examples.

When developing a computing device that transforms the input vector $\vec{x}$ into the output vector $\vec{y}$, the transformation of the function given in the high-level input language into a sequence of actions is performed at each stage of the transformation. For this purpose, a synthesizer developed in the Specialized Computer Systems Laboratory at RTU MIREA is used [4]. The output of the synthesizer comprises a text in the hardware description language, which forms registers at the stages of the pipeline and combinational logic nodes between them to perform the transformations f_1, f_2, f_3, … f_n. A similar approach is used in a number of synthesizers [5][1]. However, for the developed software product there is a possibility of synthesis control based on feedback formed by analyzing the results of component placement and tracing. In this case, the maximum value of the signal propagation delay between the stages of the pipeline determines the minimum period of the clock frequency signal. The components of this delay should be determined for the topological basis in such a way that the synthesizer can evenly distribute the signal propagation delay between the pipeline stages.

The signal propagation delay between registers of a programmable logic device (PLD) using field programmable gate array (FPGA) architecture is defined as follows[2]:

$$t = t_{\text{logic}} + t_{\text{route}}, \tag{1}$$

where t_{logic} is the propagation delay determined by the combinational elements; t_{route} is the propagation delay determined by the PLD trace circuitry.

For achieving high clock frequency and uniform distribution of total signal delay between all stages of the pipeline, the synthesizer should evaluate the components defined by the combinational elements as well as those defined by the trace circuits. Once the signal transformations have been distributed to the combinational logic nodes, the resulting register transfer level (RTL) representation of the pipeline is passed to the PLD computer-aided design (CAD) system, which performs the placement of the circuit components and tracing of the interconnects. In this case, suboptimal component placement introduces additional signal delay that violates the uniformity of delay distribution across the pipeline stages.

The stages in the development of a pipeline computing device are shown in Fig. 1.

In order to achieve a high clock frequency of the pipeline operation, it is necessary to estimate the components of the signal propagation delays and eliminate the negative effects of suboptimal mutual placement of the interconnected components on the PLD chip. While the placement optimization problem

[1] https://docs.amd.com/r/en-US/ug1399-vitis-hls/HLS-Programmers-Guide. Accessed October 10, 2024.

[2] https://docs.xilinx.com/r/en-US/ug906-vivado-design-analysis/Timing-Analysis. Accessed October 10, 2024.

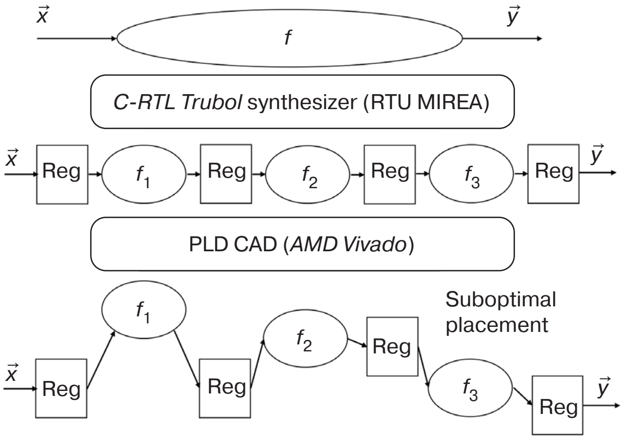


Fig. 1. The stages in the development of a pipelined computing device. The *C-RTL Trubol* synthesizer is a software developed by the authors. The synthesizer is a tool for creating a description in the electronic CAD format

can be formulated in the strict mathematical sense, it has no practical solution in an acceptable time due to the burgeoning complexity of optimization algorithms for a general formulation. In practice, the PLD-based design uses design constraints that specify the areas for placing groups of components (so-called topological design constraints). For groups of components (called P-blocks) placed in this way, optimization by CAD algorithms can be performed in a reasonable time but with suboptimal results. Technical methods for controlling design constraints are specified in the AMD UltraFast™ Design Methodology Guide for FPGAs and Systems-on-a-Chip[3]. Research into the use of topological design constraints as reflected in a number of publications and dates back to 2011 [6] is driven by the development of design tools such as *Xilinx PlanAhead* [7, 8]. At present, the use of design constraints is still being used in network packet processing [9] and digital signal processing [10].

CONTROL OF THE SYNTHESIS OPERATIONS OF THE PIPELINE FUNCTIONAL UNITS

We consider the following sequence of design of a pipelined computational structure. A set of test pipeline chains containing nodes with the appropriate logic function is created to estimate the delay determined by combinational logic. For these nodes, the pipeline is synthesized and placed, and the experimental estimate is written into the data structure passed to the *C-RTL* synthesizer. As the synthesizer is an original design, appropriate modifications are introduced to account for delays introduced by external sources.

It should be noted that the use of PLD does not require the evaluation of a wide range of possible arithmetic and logic operations. For this purpose, since bitwise operations are performed based on truth tables, while addition and subtraction operations can be carried out using special "fast carry chain" nodes, it is sufficient to estimate the delay caused by these two classes of operations.

Given an architectural pattern and certain delay parameters, the design sequence of the pipeline computing structure shown in Fig. 2 can be adopted.

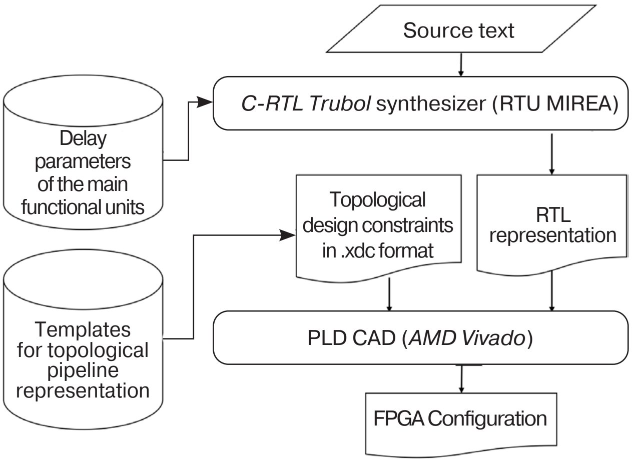

Source text
C-RTL Trubol synthesizer (RTU MIREA)
Delay parameters of the main functional units
Topological design constraints in .xdc format
RTL representation
Templates for topological pipeline representation
PLD CAD (AMD Vivado)
FPGA Configuration

Fig. 2. The sequence of the automated design of the pipeline computing structure

The input to the developed sequence is assumed to be the program source code in a problem-oriented high-level language, as well as the design constraints formed on the basis of the study of the hardware platform characteristics. The synthesis uses the delay parameters of the main functional nodes that have been preliminarily evaluated in the process of pipeline synthesis according to a predetermined scheme that explicitly distinguishes the circuit fragments for evaluation. The dedicated synthesizer generates an RTL representation of the module in the hardware description language, which is complemented by the design constraints file in .xdc or .sdc format (depending on the CAD system used). This file is created by parameterizing one of the developed templates for the topological pipeline representation.

FORMATION OF TOPOLOGICAL DESIGN CONSTRAINTS FOR PIPELINE COMPUTATION STRUCTURE

When placing a pipeline in PLD, it is necessary to specify the rules for placing its individual components. In PLD CAD systems, it is common practice to have a hierarchical (modular) placement mode whose placement algorithms use the project modules at the RTL level as localized project units to place their components

[3] https://docs.amd.com/r/en-US/ug949-vivado-design-methodology. Accessed October 10, 2024.

as compactly as possible. At the same time, the designer can accompany the project with special project constraint files that control the processes of the Implementation group, such as timing analysis and placement. The corresponding constraint language command groups deal with timing constraints and topological (area) constraints.

The command for describing design constraints within a single stage has the following form:

```
create_pb <pblock name>
< X-axis pblock coordinate >
< Y-axis pblock coordinate >
< pblock width> < pblock height>
< list of elements associated with pblock>
```

In this example, PLD on-chip boundaries with specified coordinates are set for triggers whose names match the template (Figure 3). The size of the on-chip boundaries and coordinates are precisely matched to the overall topology of the chip along with the number of registers and logic elements involved in each stage. The name template is selected based on the RTL level description format.

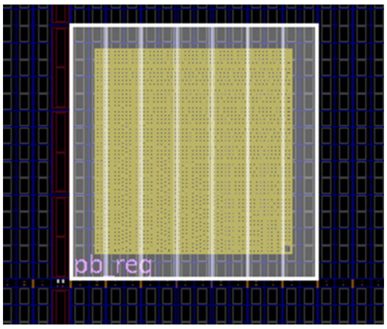


Fig. 3. The process of the P-block allocation in PLD using FPGA architecture

The study confirms the assumed utility of describing topological design constraints only for pipeline registers. This is due to the local relation of the combinational logic between the register groups and corresponding pipeline stages. By defining the pipeline stage placement regions, a compact placement of the combinational logic nodes associated with the registers is achieved while preserving the CAD capability to perform local optimizations. In this case, the manual control of separate pipeline components, including separate triggers and combinational logic nodes, would be excessively labor-intensive.

For the circuit synthesizer developed in RTU MIREA in RTL format, it is recommended to modify it by generating register names with the introduction of fragments that uniquely identify the pipeline stage to which this register belongs. Although this information is available in the internal representation of the synthesizer, it has not yet been used. Analysis of world analogs has confirmed the impossibility of identifying the pipeline stage is impossible in them; consequently, the exported RTL representation typically uses end-to-end numbering of separate circuit triggers. At the same time, the introduction of information about the pipeline stage allows the allocation of a group of registers corresponding to this stage using a regular expression having the following form:

```
*/pipeline_unit/*/*reg_Ki\_*
```

EXAMPLES OF PRACTICAL TESTING OF THE METHOD

The practical testing of the method is carried out through the implementation of several types of pipelines. For example, sequential application of the vector rotation operation is used in the CORDIC[4] algorithm [11]. Similar operations in the form of a combination of addition and shift are used in successive multiplication with accumulation. These operations can be combined in a single configurable pipeline. The schematic fragment generated in *Vivado*[5] CAD shown in Fig. 4 demonstrates the possibility of obtaining a locally optimal solution by automatic placement of pipeline components followed by their optimization.

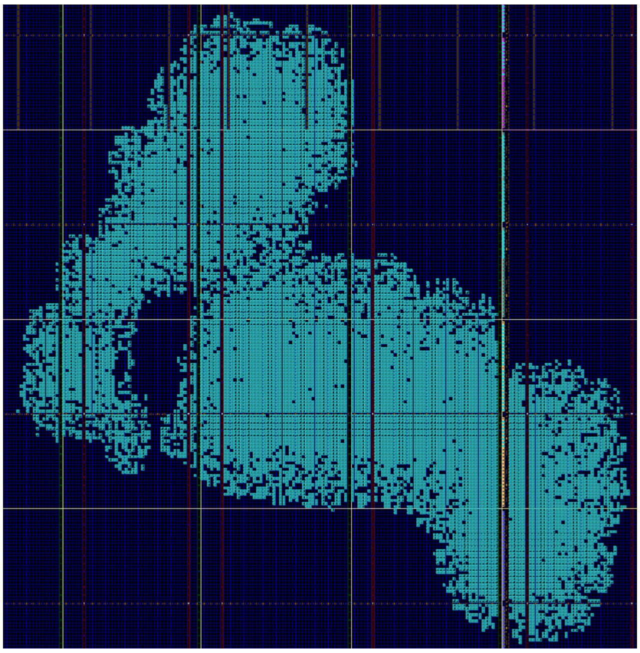


Fig. 4. Pipeline stage placement in automated mode in *Vivado* CAD

From a CAD point of view, the search for an optimal solution does not take into account the stage partitioning inherent in the pipeline architecture. Based on this

[4] CORDIC is an acronym for COordinate Rotation DIgital Computer; a “digit by digit” method.

[5] https://www.amd.com/en/products/software/adaptive-socs-and-fpgas/vivado.html. Accessed October 10, 2024.

assumption, the placement characteristics can be improved using the approach described above. The experimental results confirm the possibility of improving basic circuit performance when the block boundaries of each stage partially overlap the boundaries of adjacent blocks. The result obtained using this strategy is shown in Fig. 5.

In comparison to the standard placement (Fig. 6), a clock frequency of 1 GHz can be achieved by placement in an AMD FPGA Virtex™ UltraScale™+xcvu440_CIV-flga2892-3-e[6] (Fig. 7).

When the number of stages of the CORDIC algorithm is increased, the result of the approach becomes even

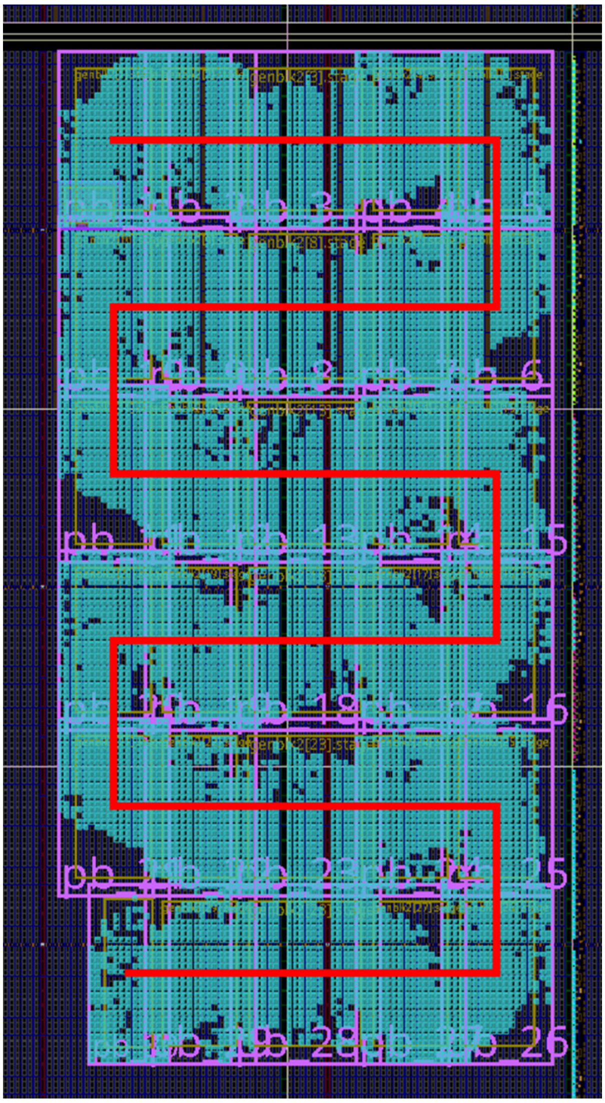
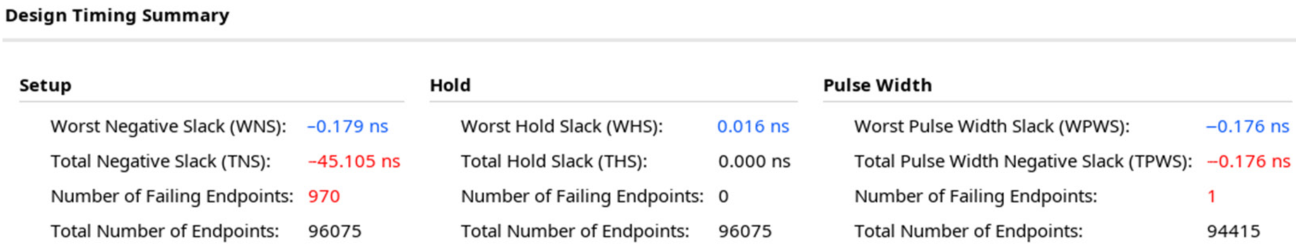


Fig. 5. Pipeline stage placement of using topological design constraints



Design Timing Summary

Setup

Worst Negative Slack (WNS): −0.179 ns

Total Negative Slack (TNS): −45.105 ns

Number of Failing Endpoints: 970

Total Number of Endpoints: 96075

Hold

Worst Hold Slack (WHS): 0.016 ns

Total Hold Slack (THS): 0.000 ns

Number of Failing Endpoints: 0

Total Number of Endpoints: 96075

Pulse Width

Worst Pulse Width Slack (WPWS): −0.176 ns

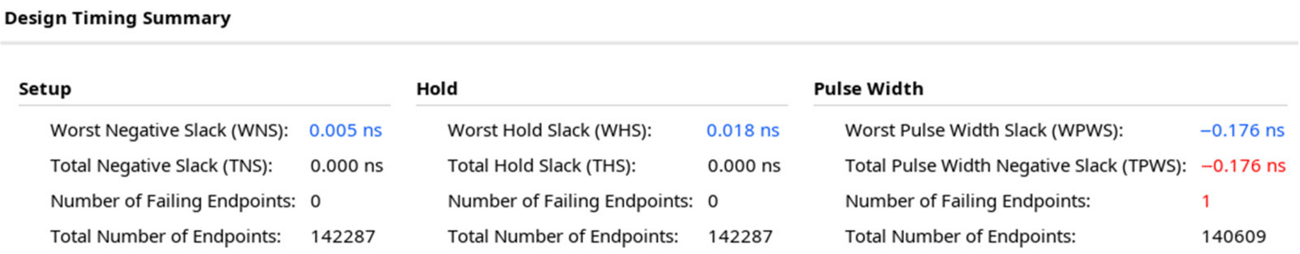
Total Pulse Width Negative Slack (TPWS): −0.176 ns

Number of Failing Endpoints: 1

Total Number of Endpoints: 94415

Fig. 6. CAD report snippet with project time characteristics for the circuit solution obtained in auto mode

[6] https://www.xilinx.com/products/boards-and-kits/1-66ql3z.html. Accessed October 10, 2024.



Design Timing Summary

Setup		Hold		Pulse Width	
Worst Negative Slack (WNS):	0.005 ns	Worst Hold Slack (WHS):	0.018 ns	Worst Pulse Width Slack (WPWS):	−0.176 ns
Total Negative Slack (TNS):	0.000 ns	Total Hold Slack (THS):	0.000 ns	Total Pulse Width Negative Slack (TPWS):	−0.176 ns
Number of Failing Endpoints:	0	Number of Failing Endpoints:	0	Number of Failing Endpoints:	1
Total Number of Endpoints:	142287	Total Number of Endpoints:	142287	Total Number of Endpoints:	140609

Fig. 7. CAD report snippet with project time characteristics for a circuit solution using topological design constraints

more pronounced. The results for 64 stages at a clock frequency of 600 MHz are shown in Figs. 8 and 9.

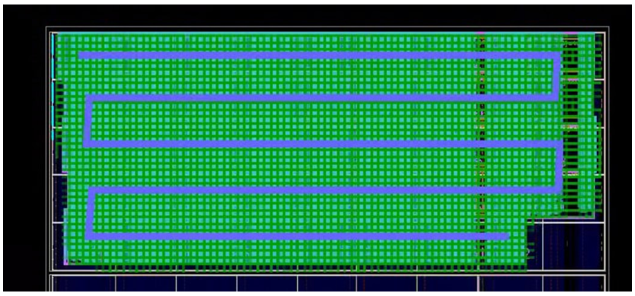


Fig. 8. Pipeline stage placement (64 stages, 600 MHz clock frequency)

A similar approach is applied to the placement of the considered circuit on the AMD FPGA Artix-7 xc7a100tcsg324-1[7] (16 stages, 400 MHz clock frequency) and AMD FPGA Kintex™ UltraScale™ xcku115_CIV-flvf1924-3-e[8] (32 stages, 850 MHz clock frequency) chips. The viability of the approach is demonstrated by the comparison of timing analysis results for the circuits under consideration (Figs. 10–13).

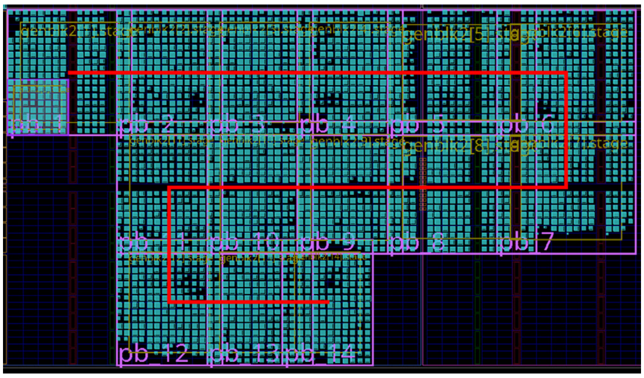
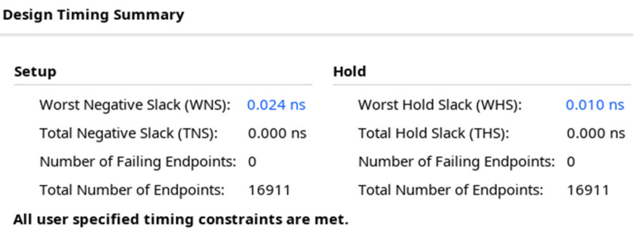


Fig. 10. Pipeline stage placement using topological design constraints (xc7a100tcsg324-1 PLD)

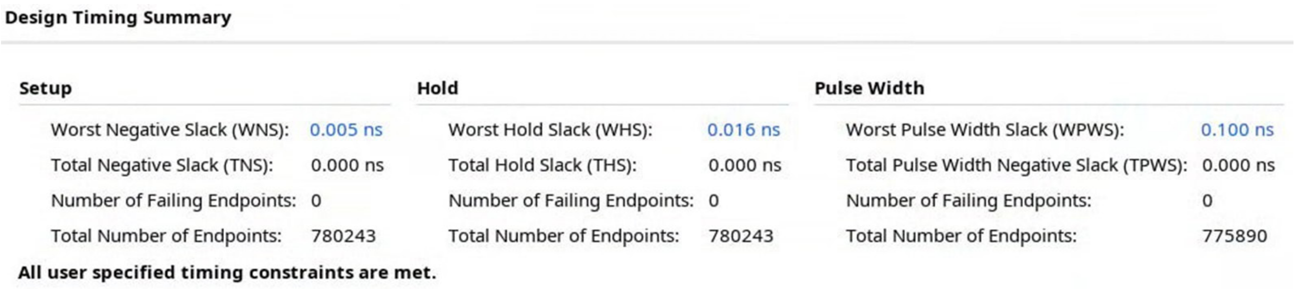


Design Timing Summary

Setup		Hold	
Worst Negative Slack (WNS):	0.024 ns	Worst Hold Slack (WHS):	0.010 ns
Total Negative Slack (TNS):	0.000 ns	Total Hold Slack (THS):	0.000 ns
Number of Failing Endpoints:	0	Number of Failing Endpoints:	0
Total Number of Endpoints:	16911	Total Number of Endpoints:	16911

All user specified timing constraints are met.

Fig. 11. CAD report snippet with project time characteristics for the obtained placement (xc7a100tcsg324-1 PLD)



Design Timing Summary

Setup		Hold		Pulse Width	
Worst Negative Slack (WNS):	0.005 ns	Worst Hold Slack (WHS):	0.016 ns	Worst Pulse Width Slack (WPWS):	0.100 ns
Total Negative Slack (TNS):	0.000 ns	Total Hold Slack (THS):	0.000 ns	Total Pulse Width Negative Slack (TPWS):	0.000 ns
Number of Failing Endpoints:	0	Number of Failing Endpoints:	0	Number of Failing Endpoints:	0
Total Number of Endpoints:	780243	Total Number of Endpoints:	780243	Total Number of Endpoints:	775890

All user specified timing constraints are met.

Fig. 9. CAD report snippet with time characteristics of the CORDIC pipeline computer project (64 stages, 600 MHz clock frequency)

[7] https://www.amd.com/en/products/adaptive-socs-and-fpgas/fpga/artix-7.html. Accessed October 10, 2024.

[8] https://www.amd.com/en/products/adaptive-socs-and-fpgas/fpga/kintex-ultrascale-plus.html. Accessed October 10, 2024.

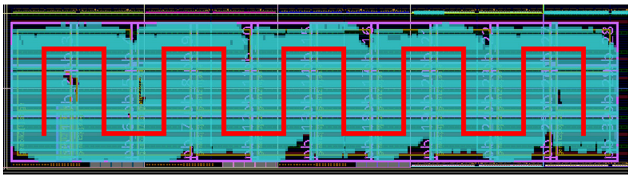
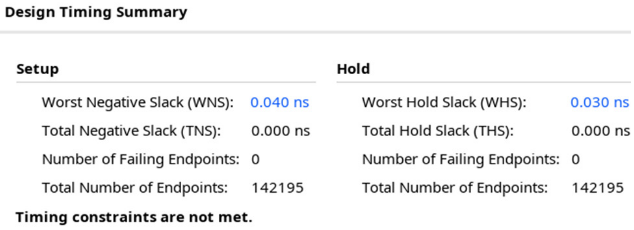


Fig. 12. Pipeline stage placement using topological design constraints (xcku115_CIV-flvf1924-3-e PLD)



```
Design Timing Summary

Setup                                          Hold
  Worst Negative Slack (WNS):    0.040 ns        Worst Hold Slack (WHS):        0.030 ns
  Total Negative Slack (TNS):    0.000 ns        Total Hold Slack (THS):        0.000 ns
  Number of Failing Endpoints:   0               Number of Failing Endpoints:   0
  Total Number of Endpoints:     142195          Total Number of Endpoints:     142195
Timing constraints are not met.
```

Fig. 13. CAD report snippet with project time characteristics for the obtained placement (xcku115_CIV-flvf1924-3-e PLD)

The results confirm the possibility of setting design constraints on registers of a separate module describing the pipeline to systematically improve the design properties of computers having a pipelined structure. The continuing interest in pipelined nodes [12] confirms the relevance of this research direction. In addition, pipelined devices can function as subsystems of computational complexes to increase their efficiency both in widespread tasks [13] and when used as accelerators of a narrow computational subclass [14, 15].

CONCLUSIONS

The presented materials describe results obtained by the Specialized Computer Systems Laboratory at RTU MIREA in the development of a design methodology for specialized pipelined computing accelerators. By focusing on a simple hardware architecture with localized connections between nodes, it is possible to develop a set of algorithms and design measures for systematically improving the topological representation properties of a computing device from its initial high-level language description. The obtained results can be adapted to other types of architectural templates to extend the nomenclature of specialized electronic component bases.

ACKNOWLEDGMENTS

The research is carried out within the framework of the State Assignment of the Ministry of Science and Higher Education of the Russian Federation (No. FSFZ-2022-0004 “Architectures of special-purpose computing units and procedures, algorithms, and tools for the design of digital computing units”).

Authors’ contribution

All authors equally contributed to the research work.

REFERENCES

1. Saidov B.B., Telezhkin V.F., Gudaev N.N., et al. Development of Equipment for Experimental Study of Digital Algorithms in Nonstationary Signal Processing Problems. *Ural Radio Engineering Journal.* 2022;6(2):186–204. https://doi.org/10.15826/urej.2022.6.2.004
2. Jasek R. SHA-1 and MD5 Cryptographic Hash Functions: Security Overview. *Communications* (*Komunikacie*). 2015;17(1):73–80.
3. Carrión D.S., Prohaska V., Diez O. Exploration of TPUs for AI Applications. In: Daimi K., Al Sadoon A. (Eds.). *Proceedings of the Second International Conference on Advances in Computing Research* (*ACR’24*). ACR 2024. Lecture Notes in Networks and Systems. Springer; 2024. V. 956. P. 559. https://doi.org/10.1007/978-3-031-56950-0_47
4. Tarasov I.E., Sovietov P.N., Lulyava D.V., Mirzoyan D.I. Method for designing specialized computing systems based on hardware and software co-optimization. *Russian Technological Journal.* 2024;12(3):37–45. https://doi.org/10.32362/2500-316X-2024-12-3-37-45
5. Alekhin V.A. Designing Electronic Systems Using SystemC and SystemC–AMS. *Russian Technological Journal.* 2020;8(4):79–95 (in Russ.). https://doi.org/10.32362/2500-316X-2020-8-4-79-95
6. Pham-Quoc C., Dinh-Duc A.-V. Automatic generation of area constraints for FPGA implementation. In: *2011 IEEE 3rd International Conference on Communication Software and Networks* (*ICCSN*). 2011. P. 469–472. https://doi.org/10.1109/ICCSN.2011.6014937
7. Li K., Lei L., Guang Q., Shi J.-Y., Hao Y. Improving the performance of an SOC design for network processing based on FPGA with PlanAhead. In: 2011 *International Conference on Electronics, Communications and Control* (*ICECC*). 2011. P. 297–300. https://doi.org/10.1109/ICECC.2011.6066640
8. Sarker A.L.Md., Lee M.H. Synthesis of VHDL code for FPGA design flow using Xilinx PlanAhead tool. In: *2012 International Conference on Education and e-Learning Innovations* (*ICEELI*). 2012. https://doi.org/10.1109/ICEELI.2012.6360614

9. Song X., Lu R., Guo Z. High-Performance Reconfigurable Pipeline Implementation for FPGA-Based SmartNIC. *Micromachines.* 2024;15(4):449. https://doi.org/10.3390/mi15040449
10. Anderson T., Wheeler T.J. An FPGA-based hardware accelerator supporting sensitive sequence homology filtering with profile hidden Markov models. *BMC Bioinformatics.* 2024;25:247. https://doi.org/10.1186/s12859-024-05879-3
11. Tarasov I.E., Sovetov P.N. *Device for Calculating Transcendental Functions and Multiplying Binary Numbers*: Pat. 222880 U1 RF. Publ. 22.01.2024 (in Russ.).
12. Oishi R., Kadomoto J., Irie H., Sakai S. FPGA-based Garbling Accelerator with Parallel Pipeline Processing. *IEICE Trans. Inform. Syst.* 2023;E106.D(12):1988–1996. https://doi.org/10.1587/transinf.2023PAP0002
13. Nurvitadhi E., Sheffield D., Sim J., et al. Accelerating Binarized Neural Networks: Comparison of FPGA, CPU, GPU, and ASIC. In: *2016 International Conference on Field-Programmable Technology* (*FPT*). 2016. P. 77–84. https://doi.org/10.1109/FPT.2016.7929192
14. Hennessy J.L., Patterson D.A. A new golden age for computer architecture: Domain-specific hardware/software co-design, enhanced security, open instruction sets, and agile chip development. In: *Proceedings of the 2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture* (*ISCA*). 2018. P. 27–29. https://doi.org/10.1109/ISCA.2018.00011
15. Hennessy J.L., Patterson D.A. *Computer Architecture*: *A Quantitative Approach:* 6th ed. The Morgan Kaufmann Series in Computer Architecture and Design. 2017. 936 p.

СПИСОК ЛИТЕРАТУРЫ

1. Saidov B.B., Telezhkin V.F., Gudaev N.N., et al. Development of Equipment for Experimental Study of Digital Algorithms in Nonstationary Signal Processing Problems. *Ural Radio Engineering Journal.* 2022;6(2):186–204. https://doi.org/10.15826/urej.2022.6.2.004
2. Jasek R. SHA-1 and MD5 Cryptographic Hash Functions: Security Overview. *Communications* (*Komunikacie*). 2015;17(1):73–80.
3. Carrión D.S., Prohaska V., Diez O. Exploration of TPUs for AI Applications. In: Daimi K., Al Sadoon A. (Eds.). *Proceedings of the Second International Conference on Advances in Computing Research* (*ACR'24*). ACR 2024. Lecture Notes in Networks and Systems. Springer; 2024. V. 956. P. 559. https://doi.org/10.1007/978-3-031-56950-0_47
4. Тарасов И.Е., Советов П.Н., Люлява Д.В., Мирзоян Д.И. Методика проектирования специализированных вычислительных систем на основе совместной оптимизации аппаратного и программного обеспечения. *Russian Technological Journal.* 2024;12(3):37–45 https://doi.org/10.32362/2500-316X-2024-12-3-37-45
5. Алехин В.А. Проектирование электронных систем с использованием SystemC и SystemC–AMS. *Russian Technological Journal.* 2020;8(4):79–95. https://doi.org/10.32362/2500-316X-2020-8-4-79-95
6. Pham-Quoc C., Dinh-Duc A.-V. Automatic generation of area constraints for FPGA implementation. In: *2011 IEEE 3rd International Conference on Communication Software and Networks* (*ICCSN*). 2011. P. 469–472. https://doi.org/10.1109/ICCSN.2011.6014937
7. Li K., Lei L., Guang Q., Shi J.-Y., Hao Y. Improving the performance of an SOC design for network processing based on FPGA with PlanAhead. In: 2011 *International Conference on Electronics, Communications and Control* (*ICECC*). 2011. P. 297–300. https://doi.org/10.1109/ICECC.2011.6066640
8. Sarker A.L. Md, Lee M.H. Synthesis of VHDL code for FPGA design flow using Xilinx PlanAhead tool. In: *2012 International Conference on Education and e-Learning Innovations* (*ICEELI*). 2012. https://doi.org/10.1109/ICEELI.2012.6360614
9. Song X., Lu R., Guo Z. High-Performance Reconfigurable Pipeline Implementation for FPGA-Based SmartNIC. *Micromachines.* 2024;15(4):449. https://doi.org/10.3390/mi15040449
10. Anderson T., Wheeler T.J. An FPGA-based hardware accelerator supporting sensitive sequence homology filtering with profile hidden Markov models. *BMC Bioinformatics.* 2024;25:247. https://doi.org/10.1186/s12859-024-05879-3
11. Тарасов И.Е., Советов П.Н. *Устройство для вычисления трансцендентных функций и умножения двоичных чисел*: пат. 222880 U1 РФ. Заявка № 2023131099; заявл. 28.11.2023; опубл. 22.01.2024. Бюл. № 3.
12. Oishi R., Kadomoto J., Irie H., Sakai S. FPGA-based Garbling Accelerator with Parallel Pipeline Processing. *IEICE Trans. Inform. Syst.* 2023;E106.D(12):1988–1996. https://doi.org/10.1587/transinf.2023PAP0002
13. Nurvitadhi E., Sheffield D., Sim J., et al. Accelerating Binarized Neural Networks: Comparison of FPGA, CPU, GPU, and ASIC. In: *2016 International Conference on Field-Programmable Technology* (*FPT*). 2016. P. 77–84. https://doi.org/10.1109/FPT.2016.7929192
14. Hennessy J.L., Patterson D.A. A new golden age for computer architecture: Domain-specific hardware/software co-design, enhanced security, open instruction sets, and agile chip development. In: *Proceedings of the 2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture* (*ISCA*). 2018. P. 27–29. https://doi.org/10.1109/ISCA.2018.00011
15. Hennessy J.L., Patterson D.A. *Computer Architecture*: *A Quantitative Approach:* 6th ed. The Morgan Kaufmann Series in Computer Architecture and Design. 2017. 936 p.

About the Authors

Ilya E. Tarasov, Dr. Sci. (Eng.), Associated Professor, Head of the Laboratory of Specialized Computing Systems, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: tarasov_i@mirea.ru. Scopus Author ID 57213354150, RSCI SPIN-code 4628-7514, http://orcid.org/0000-0001-6456-4794

Peter N. Sovietov, Cand. Sci. (Eng.), Senior Researcher, Laboratory of Specialized Computing Systems, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: peter.sovietov@gmail.com. Scopus Author ID 57221375427, RSCI SPIN-code 9999-1460, http://orcid.org/0000-0002-1039-2429

Daniil V. Lulyava, Junior Researcher, Laboratory of Specialized Computing Systems, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: lyulyava@mirea.ru. Scopus Author ID 58811698000, RSCI SPIN-code 1882-0989, http://orcid.org/0009-0009-9623-7777

Nikita A. Duksin, Engineer, Laboratory of Specialized Computing Systems, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: duksin@mirea.ru. RSCI SPIN-code 1082-8956, Scopus Author ID 58811361100, https://orcid.org/0009-0009-0014-7065

Об авторах

Тарасов Илья Евгеньевич, д.т.н., доцент, заведующий лабораторией специализированных вычислительных систем, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: tarasov_i@mirea.ru. Scopus Author ID 57213354150, SPIN-код РИНЦ 4628-7514, http://orcid.org/0000-0001-6456-4794

Советов Петр Николаевич, к.т.н., старший научный сотрудник, лаборатория специализированных вычислительных систем, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: peter.sovietov@gmail.com. Scopus Author ID 57221375427, SPIN-код РИНЦ 9999-1460, http://orcid.org/0000-0002-1039-2429

Люлява Даниил Вячеславович, младший научный сотрудник, лаборатория специализированных вычислительных систем, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: lyulyava@mirea.ru. Scopus Author ID 58811698000, SPIN-код РИНЦ 1882-0989, http://orcid.org/0009-0009-9623-7777

Дуксин Никита Александрович, инженер, лаборатория специализированных вычислительных систем, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: duksin@mirea.ru. SPIN-код РИНЦ 1082-8956, Scopus Author ID 58811361100, https://orcid.org/0009-0009-0014-7065

Translated from Russian into English by K. Nazarov
Edited for English language and spelling by Thomas A. Beavitt

UDC 681.515
https://doi.org/10.32362/2500-316X-2025-13-3-54-62
EDN SHAEZM

CC BY

RESEARCH ARTICLE

Analysis and synthesis of intelligent automatic control systems with type-1 fuzzy regulator

Yuri A. Bykovtsev @,
Valery M. Lokhin

MIREA – Russian Technological University, Moscow, 119454 Russia
@ Corresponding author, e-mail: bykovcev@mirea.ru

• **Submitted:** 20.11.2024 • **Revised:** 13.02.2025 • **Accepted:** 21.03.2025

Abstract
Objectives. The active development of intelligent automatic control systems, which is associated with increasing requirements to the quality and accuracy of control systems of modern technical systems, requires the development of new approaches to their analysis and synthesis. A promising class of intelligent control devices is based on regulators that use fuzzy-logic inference technology. The purpose of this work is to develop a method for the complex synthesis of type-1 fuzzy regulator parameters on the basis of the Yakubovich circle criterion.
Methods. The proposed methodology is based on a consideration of fuzzy regulators in terms of the corresponding nonlinear transformation that support the use of methods derived from the theory of nonlinear automatic control systems. Analogs of the degrees of stability and oscillation are used as quality indicators. The synthesis of the parameters of the nonlinear transformation can be reduced to determining sufficient regions of absolute stability of the system with the shifted and extended Nyquist plot obtained using the Yakubovich circle stability criterion.
Results. In accordance with the theory of fuzzy sets and algorithms of fuzzy logical inference described by Takagi–Sugeno, the possibility of one-to-one correspondence of the nonlinear transformation and the parameters of an appropriately arranged knowledge base of the fuzzy controller is shown. A procedure proposed for synthesizing the parameters of the type-1 fuzzy regulator is aimed at ensuring complex requirements for the quality of the control system according to the degree of stability, the degree of oscillation, and steady-state mode accuracy. The effectiveness of the proposed technique, which guarantees the absolute stability not only of the equilibrium position but also of the processes, is confirmed by the results of model experiments.
Conclusions. The paper proposes a convenient engineering technique for determining the parameters of an intelligent controller constructed using fuzzy logic inference technology based on methods informed by automatic control theory. The convenience of using such indirect quality indicators as the degree of stability, the degree of oscillation, and accuracy in steady-state mode, is demonstrated. These indicators are explicable for developers of applied control systems.

Keywords: intelligent control system, fuzzy logic inference, fuzzy controller, Takagi–Sugeno model, absolute stability of processes

© Y.A. Bykovtsev, V.M. Lokhin, 2025

For citation: Bykovtsev Y.A., Lokhin V.M. Analysis and synthesis of intelligent automatic control systems with type-1 fuzzy regulator. *Russian Technological Journal.* 2025;13(3):54–62. https://doi.org/10.32362/2500-316X-2025-13-3-54-62, https://www.elibrary.ru/SHAEZM

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Анализ и синтез интеллектуальных систем автоматического управления с нечетким регулятором I рода

Ю.А. Быковцев @,
В.М. Лохин

МИРЭА – Российский технологический университет, Москва, 119454 Россия
@ *Автор для переписки, e-mail: bykovcev@mirea.ru*

• **Поступила:** 20.11.2024 • **Доработана:** 13.02.2025 • **Принята к опубликованию:** 21.03.2025

Резюме

Цели. Активное развитие интеллектуальных систем автоматического управления, связанное с повышением требований к качеству и точности систем управления современных технических систем, требует разработки новых подходов к их анализу и синтезу. Одним из перспективных классов интеллектуальных управляющих устройств выступают регуляторы, построенные на базе технологии нечеткого логического вывода. Целью настоящей работы является разработка методики комплексного синтеза параметров нечеткого регулятора I рода на основе кругового критерия Якубовича.

Методы. В основу предлагаемой методики положено рассмотрение нечеткого регулятора с позиции соответствующего нелинейного преобразования, что позволяет использовать методы теории нелинейных систем автоматического управления. В качестве показателей качества в работе используются аналоги понятий «степень устойчивости» и «степень колебательности». Синтез параметров нелинейного преобразования сводится к определению достаточных областей абсолютной устойчивости системы со смещенной и расширенной амплитудно-фазовыми частотными характеристиками, полученных с помощью кругового критерия устойчивости Якубовича.

Результаты. В соответствии с теорией нечетких множеств и алгоритмом нечеткого логического вывода Такаги – Сугено показана возможность взаимно-однозначного соответствия нелинейного преобразования и параметров базы знаний нечеткого регулятора при соответствующей организации последней. В работе предложена процедура синтеза параметров нечеткого регулятора I рода, нацеленная на обеспечение комплексных требований к качеству системы управления по «степени устойчивости», «степени колебательности» и точности в установившемся режиме. Предложенная методика также гарантирует абсолютную устойчивость не только положения равновесия, но и процессов, а ее эффективность подтверждена результатами модельных экспериментов.

Выводы. В работе предложена удобная инженерная методика настройки параметров интеллектуального регулятора, построенная по технологии нечеткого логического вывода на основе методов теории автоматического управления. Показано удобство применения таких косвенных показателей качества, как «степень устойчивости», «степень колебательности» и точность в установившемся режиме.

Ключевые слова: интеллектуальная система управления, нечеткий логический вывод, нечеткий регулятор, модель Такаги – Сугено, абсолютная устойчивость процессов

Для цитирования: Быковцев Ю.А., Лохин В.М. Анализ и синтез интеллектуальных систем автоматического управления с нечетким регулятором I рода. *Russian Technological Journal*. 2025;13(3):54−62. https://doi.org/10.32362/2500-316X-2025-13-3-54-62, https://www.elibrary.ru/SHAEZM

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

Intelligent technologies have been increasingly applied in various fields of activity over the last two or three decades. One such field is automatic control systems (ACS) [1], for which a new generation of controllers based on intelligent technologies (expert systems, neural-like networks, associative memory or fuzzy logic) is being developed. These intelligent controllers not only provide high quality ACS performance, but are also capable of adaptation to various uncertainties affecting the system.

Among these intellectual technologies, fuzzy inference or fuzzy logic is the most widely used for both objective and subjective reasons [2–4]. In robotics, fuzzy logic is already commonly applied in the control systems of autonomous and semi-automatic robots of various types, in the control systems of complex technological equipment, as well as at all hierarchical levels of intelligent control systems (strategic, tactical, and executive). This is largely due to the possibility of using a fuzzy inference system (FIS) to construct control models even for complex objects at the level of logical-linguistic reasoning.

However, this novel formalism is not entirely compatible with current automatic control theory (ACT). In particular, a serious problem has arisen in connection with the creation of new approaches to solving the stability and quality evaluation problems pertaining to a new class of automatic control systems. This problem has been solved quite actively in the last two decades. A comprehensive generalization of the work is given in the monograph by Pegat [5]. The concept proposed at the beginning of the present century in the RTU MIREA by Makarov et al. has turned out to be very promising [1]. Summarizing the long-term research experience in the field of fuzzy ACS taking into account the results of the studies presented in [1–6], it can be stated that:

1. Fuzzy inference allows the synthesis of logical-linguistic control models for complex objects.
2. Despite the apparent complexity of the fuzzy inference formalism, it has been established that the fuzzy regulators (FR) based on this technology are essentially nonlinear. This means that they implement a nonlinear transformation, whose parameters can change slightly when the fuzzy inference technology is modified (Mamdani, Sugeno, etc.).
3. The nature of the nonlinear transformation in FR unambiguously determines the parameters of the input logical-linguistic variables.

The representation of FR as a nonlinear ACS element provides a broad perspective for incorporating traditional approaches to nonlinear systems adopted in ACT and modified taking the specific nature of nonlinear transformations in the study of intelligent ACS into account.

SPECIFIC FEATURES OF ACS ANALYSIS AND SYNTHESIS WITH FR

We consider a fuzzy logic system having an input (E) and an output (U) linguistic variable with reasoning domains on $\mathrm{X}_E \subseteq \mathbb{R}$ and $\mathrm{Y}_U \subseteq \mathbb{R}$, for which the corresponding term sets T_E and T_U are given. Each value of a linguistic variable from the underlying term set is given by the normal fuzzy sets $\mathrm{A}_i^E = \{(\mu_\mathrm{A}(e), e) \mid e \in \mathrm{X}_E\}$ and $\mathrm{A}_i^U = \{(\mu_\mathrm{A}(u), u) \mid u \in \mathrm{X}_U\}$. Fuzzy inference models currently in active use include the Mamdani, Larsen, Takagi–Sugeno, and Tsukamoto models, which have their relative advantages and disadvantages [7]; regardless of the model type, the resulting fuzzy transformation can be represented as a certain nonlinear mapping f: $\mathrm{X}_E \to \mathrm{Y}_E$.

Nevertheless, some general principles of FR design have been formulated by most developers of fuzzy control systems, namely:

1. The number of fuzzy sets in the underlying term sets: 5–7.
2. The term set should contain at least one fuzzy set defined by the membership functions (MF) of classes L and γ to limit the control value. This is related to the ACS physical characteristics.
3. The symmetry of the MF position with respect to the central MF to ensure control symmetry when the system state deviates from equilibrium.

The division of FR into type 1 and type 2 proposed in [1] depends on the processing method of the input variables. This paper considers the ACS using the most popular type-1 FR based on the Takagi–Sugeno fuzzy inference model to be the most promising. This is primarily due to the lightweight defuzzification procedure representing the weighted mean calculation, which requires significantly less hardware resources and controller processor time compared to other methods. Additionally, the mapping implemented by the fuzzy model involving a piecewise linear function, which is dependent on the appropriate arrangement of the knowledge base, greatly simplifies both the analysis and synthesis of fuzzy ACS. The latter factor will be discussed in detail.

An excerpt of the fuzzy system is shown in Fig. 1, where two fuzzy sets A_{i-1}^{E} and A_{i}^{E} are defined by the class t MF triples $\{a_{i-1}, b_{i-1}, c_{i-1}\}$ and $\{a_i, b_i, c_i\}$, respectively, on some reasoning interval.

We will obtain the expression for the mapping f in the range of the input actions $[b_{i-1}, b_i]$. Let the rule base contain two condition-action rules having the following form:

1. If E is A_{i-1}^{E} then $U = u_{i-1}^{*}$,
2. If E is A_{i}^{E} then $U = u_{i}^{*}$,

where $u_{i-1}^{*} \equiv \text{const}$, $u_{i}^{*} \equiv \text{const}$.

In this case, the degree values of the linguistic variable E belonging to the fuzzy sets A_{i-1}^{E} and A_{i}^{E} during fuzzification are defined as follows:

$$\mu_{i-1}(e) = \frac{c_{i-1} - e}{c_{i-1} - b_{i-1}}, \quad \mu_i(e) = \frac{e - a_i}{b_i - a_i}. \tag{1}$$

According to the Takagi–Sugeno fuzzy inference procedure and the adopted constraints on MF parameter values, the output variable is defined as follows:

$$\begin{aligned} u_i(e) &= \frac{\mu_{i-1}u_{i-1}^{*} + \mu_i u_i^{*}}{\mu_{i-1} + \mu_i} = \\ &= \left(\frac{b_i - e}{b_i - b_{i-1}}u_{i-1}^{*} + \frac{e - b_{i-1}}{b_i - b_{i-1}}u_i^{*}\right) \div \left(\frac{b_i - e}{b_i - b_{i-1}} + \frac{e - b_{i-1}}{b_i - b_{i-1}}\right) = \\ &= \frac{u_i^{*} - u_{i-1}^{*}}{b_i - b_{i-1}}e + \frac{b_i u_{i-1}^{*} - b_{i-1}u_i^{*}}{b_i - b_{i-1}}. \end{aligned} \tag{2}$$

Thus, the mapping f is a linear function on the interval $[b_{i-1}; b_i]$, whose definition area is given by the location of the vertices of adjacent MFs, while its range is given by the value of the rule base conclusions.

Using the results of the above analysis, the fuzzy ACS based on the above principles can be represented as a nonlinear system, whose FR static characteristic is piecewise linear (Fig. 2). Thus, it is possible to carry out a comprehensive analysis of the dynamics of the fuzzy ACS according to the theory of nonlinear systems with regard to the nonlinear transformation characteristics.

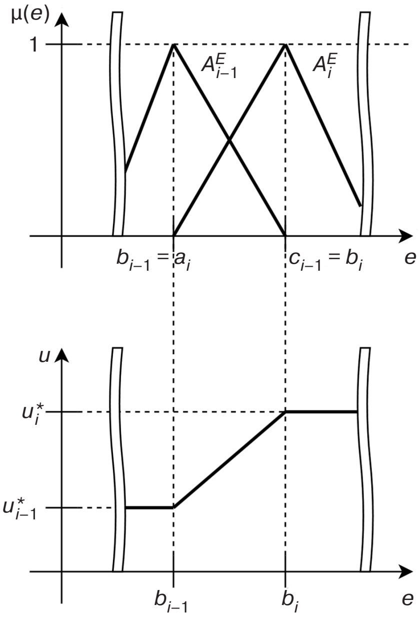


Fig. 1. Relationship between the MF parameters and the type of the nonlinear mapping

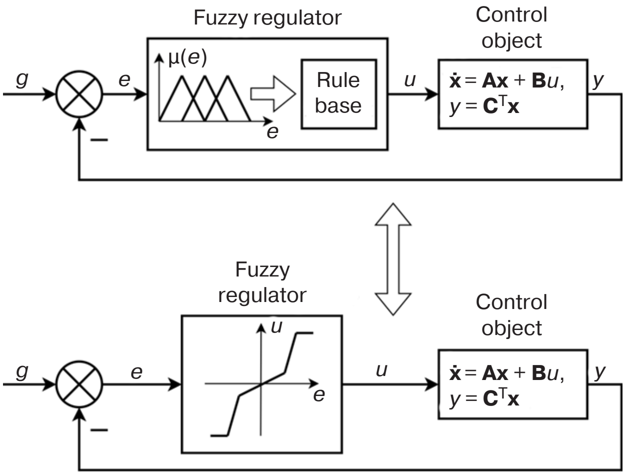

Fuzzy regulator
Control object
g
e
μ(e)
Rule base
e
u
$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{B}u$,
$y = \mathbf{C}^{\mathrm{T}}\mathbf{x}$
y
–
Fuzzy regulator
Control object
g
e
u
e
u
$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{B}u$,
$y = \mathbf{C}^{\mathrm{T}}\mathbf{x}$
y
–

Fig. 2. Conversion of an intelligent ACS with FR into a nonlinear ACS. Here, g is the master control; u is the control action; y is the control object output; $\mathbf{x}$ is the state vector; $\mathbf{A}$ is the system matrix; $\mathbf{B}$ is the control matrix; $\mathbf{C}$ is the output matrix

So far, several approaches have been developed for analyzing the stability of systems with FR. As noted in [5], despite the intensive development of new methods, the second Lyapunov method [8, 9] and the absolute stability criterion [10, 11] continue to be the most widely used. For single-input-single-output (SISO) systems, the Popov criterion and the circular criterion are recommended. For multiple-input-multiple-output (MIMO) systems, the most appropriate methods are those based on the hyperstability criterion due to providing a rigorous mathematical basis for stability evaluation.

The lack of convenient engineering methods to evaluate the qualitative parameters of these systems, such as speed and overshoot, is evident in a sufficient number of studies on the stability of fuzzy systems. This approach to the problem is more constructive due to the stability problem being automatically solved by achieving the required level of ACS quality. However, as shown by the analysis of existing studies, this problem has been insufficiently studied and remains to be worked out.

In addition to analytical methods for studying fuzzy ACS, methods based on numerical optimization algorithms including genetic algorithms [12, 13], methods related to particle swarm behavior [13, 14], gradient descent [15, 16], and others, are currently gaining wide popularity. Although these methods are powerful and capable of automatically determining the FR parameters, their use in practical engineering is exacerbated by a number of factors. Firstly, in order to use these techniques effectively, the quality function should be precisely defined, which can be a challenge. In addition, the algorithms do not provide any guidance for further adjustment of the FR parameters following their calculation on the basis of the quality criteria.

In view of the obvious potential of fuzzy systems and the relevance of FR implementation in a wide range of control systems for industrial and special purposes, the above analysis demonstrates a need to create new approaches to the study of the dynamics of such systems based on the traditional methods adopted in ACT. Based on the proposed concept, in which fuzzy ACS is considered as a nonlinear system, a suitable platform can be created for this purpose.

It is convenient to take analogs of known quality indicators for linear ACS as initial quality indicators, in particular, the stability degree as an indicator of transient damping rate and the oscillation degree as an indicator of oscillation damping. By considering FR in terms of its static characteristics, it is possible to adapt known methods for analyzing and synthesizing nonlinear ACS (in particular the Yakubovich circle stability criterion). In [17], it is shown that this criterion can be applied to the shifted amplitude-phase-frequency response (APFR) of the linear part to determine the sectional constraints on the nonlinear transformation that guarantee a certain degree of stability for the fuzzy ACS, as well as the absolute stability of the equilibrium position and processes. In this respect, an extension of this method to include oscillation degree and fuzzy ACS accuracy requirements seems promising.

ALGORITHM FOR SETTING I TYPE FR PARAMETERS BASED ON INDIRECT PERFORMANCE INDICATORS

The fuzzy system model, analogous to describing nonlinear systems adopted in ACT, can be represented as follows:

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}u,\\ u &= f(y),\\ y &= \mathbf{C}^{\mathrm{T}}\mathbf{x},\end{aligned} \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^n$, $u \in \mathbb{R}^1$, $f(y)$ is a scalar function which is the FR static input-output characteristic and belongs to the class $(K_1; K_2)$ and thus satisfies the equation [2], as follows:

$$K_1 \le \frac{df(y)}{dy} \le K_2. \tag{4}$$

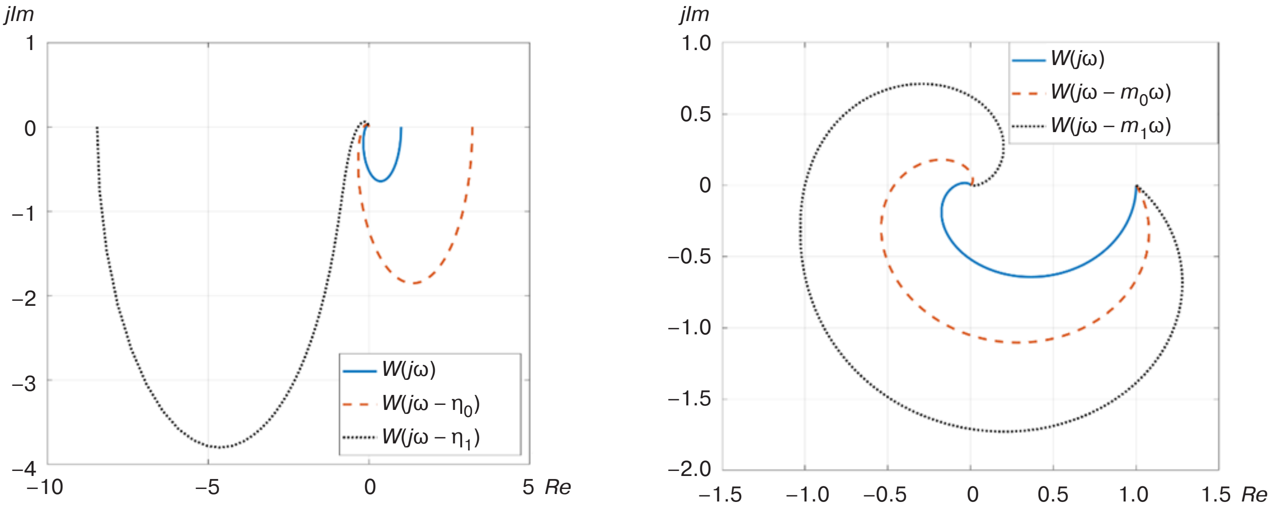
Here the task consists in synthesizing the appropriate FR knowledge base for providing the absolute process stability in fuzzy ACS and the qualitative characteristics of the transient process such as the degree of stability, the degree of oscillation, and accuracy.

For further study, it would be convenient to use the shifted $\bar{W}(j\omega - \eta)$ and extended $\hat{W}(j\omega - m\omega)$ APFR of the linear part, where $W(j\omega) = \mathbf{C}^{\mathrm{T}}(j\omega\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$, η is the analog of the stability degree and m is the analog of the oscillation degree. The family of shifted ($\eta_1 > \eta_0 > 0$) and extended ($m_1 > m_0 > 0$) APFRs for 3rd order linear parts are shown in Figs. 3a and 3b, respectively. The characteristic change in the shifted APFR of the linear part at $\eta = \eta_0$ is due to the transition of one of the poles $W(s)$ to the right complex half-plane.

Let the fuzzy control system be required to provide a fast performance not exceeding the decay time of the exponent $e^{-n_0 t}$, with n_0 chosen such that the degree of irregularity (r) of the shifted linear part $\bar{W}(j\omega - \eta)$ is equal to one.

Using the modification of the circle criterion proposed in [17], the following sufficient condition can be formulated: A fuzzy system, which is absolutely stable in class $(K_1^{\mathrm{S}}; K_2^{\mathrm{S}})$, has a degree of stability not less than η provided that the shifted APFR for the linear part of $\bar{W}(j\omega - \eta)$ covers the circle with the center on the real axis at point $-\dfrac{1}{2}\left(\dfrac{1}{K_1^{\mathrm{S}}} + \dfrac{1}{K_2^{\mathrm{S}}}\right)$, passing through points $-\dfrac{1}{K_1^{\mathrm{S}}}$ and $-\dfrac{1}{K_2^{\mathrm{S}}}$, which also belong to the real axis, r times.

The extended AFC should be outside the above circle to ensure the required degree of oscillation m. Further, if the required degree of oscillation is provided in section $(K_1^{\mathrm{O}}; K_2^{\mathrm{O}})$, then both quality requirements are met in section $\left\{\max(K_1^{\mathrm{O}}; K_1^{\mathrm{S}}); \min(K_2^{\mathrm{O}}; K_2^{\mathrm{S}})\right\}$, given the results obtained based on the specified stability. Thus, it is possible to determine the parameters K_1 and K_2



(a) (b)

Рис. 3. Family of shifted (a) and extended (b) APFRs of the linear part

of the nonlinear transformation and the corresponding parameters of the FR settings by considering the fuzzy ACS as nonlinear and applying the Yakubovich circle criterion.

Finally, we consider the problem of ensuring fuzzy ACS accuracy. Since the stationary control error of the system is determined by the part of the static characteristic close to the equilibrium, the choice of the gain factor K_1^{a} is determined by the requirement of the desired accuracy in the whole range of disturbances *f*. It is shown in [18] that for static linear parts, the stationary error in the range of disturbances $f \le \dfrac{b_1(1+K_1^{\mathrm{a}}K_{\mathrm{lp}})}{K_{\mathrm{lp}}}$ is defined as follows:

$$e_{\mathrm{stnr}} = \frac{fK_{\mathrm{lp}}}{1+K_{\mathrm{lp}}K_1^{\mathrm{a}}}, \qquad (6)$$

where b_1 is the right boundary of the static characteristic section with the gain factor K_1^{a}, while $K_{\mathrm{lp}} = \lim\limits_{\omega \to 0} W(j\omega)$.

It follows from (6) that if the maximum disturbance value $f_{\mathrm{M}} = \sup(f)$ is known and the permissible control error value e_{p} is set, the required gain factor K_1^{a} and the section boundary b_1 are defined as follows:

$$K_1^{\mathrm{a}} = \frac{f_{\mathrm{M}}K_{\mathrm{lp}} - e_{\mathrm{p}}}{K_{\mathrm{lp}}e_{\mathrm{p}}}, \qquad (7)$$

$$b_1 \ge \frac{f_{\mathrm{M}}K_{\mathrm{lp}}}{1+K_{\mathrm{lp}}K_1^{\mathrm{a}}}. \qquad (8)$$

Thus, the condition of ensuring the required quality indicators of transients and steady-state error is met in the section $(\max\{\max\{K_1^{\mathrm{O}}, K_1^{\mathrm{S}}\}, K_1^{\mathrm{a}}\}; \min\{K_2^{\mathrm{O}}, K_2^{\mathrm{S}}\})$ that ensures the absolute stability of the fuzzy ACS.

We consider an example of the synthesis of the I type FR parameters for the problem of stabilizing the fuzzy ACS equilibrium with a linear part of the 3rd order with a frequency response *W*(*j*ω). In accordance with the methodology discussed above, the necessary constructions are shown in Fig. 4. From these it follows that:

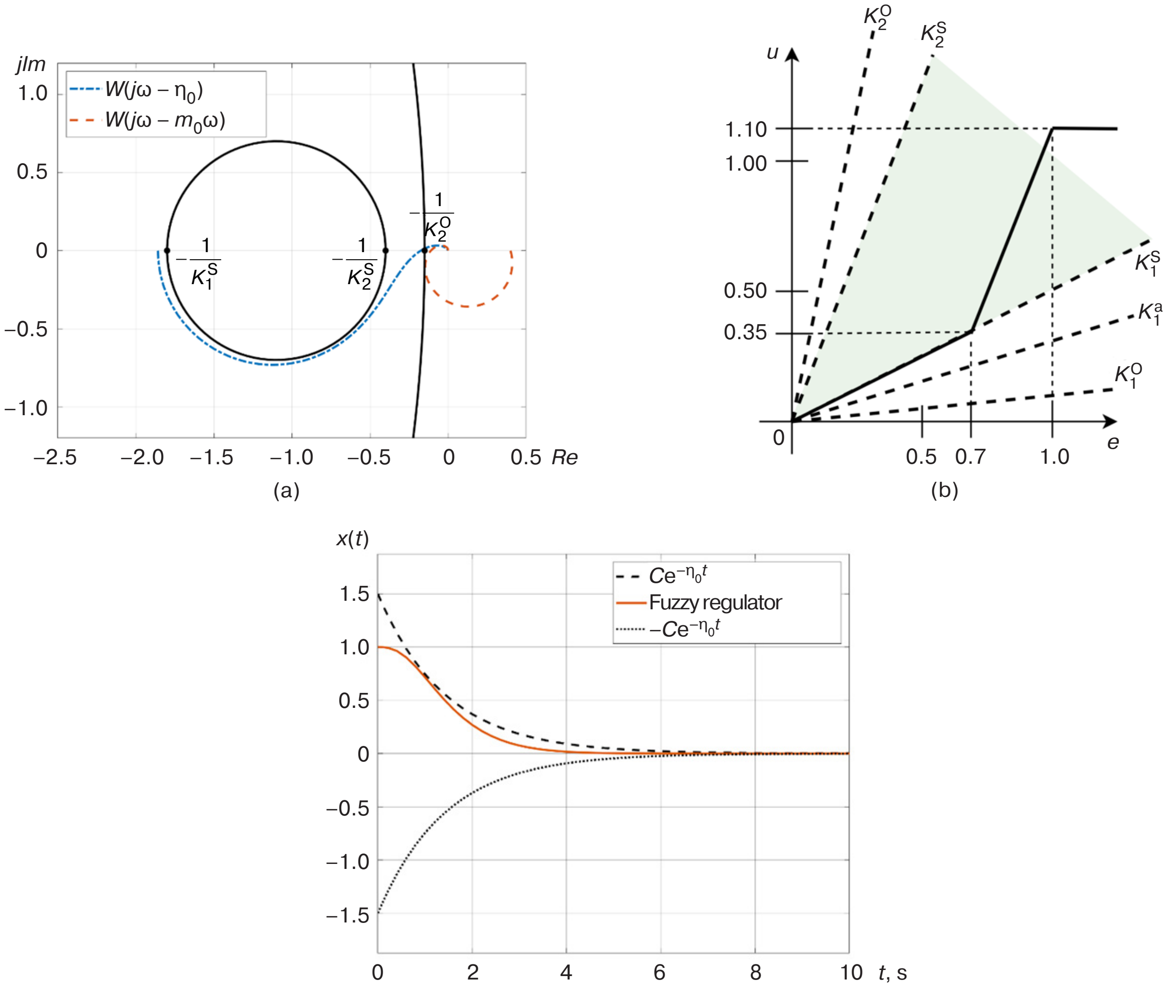
- the required value η_0 is ensured if the FR characteristic is in section (0.6; 2.5);
- the required value m_0 is provided in section (0.05; 6.1).

If the steady-state error requirement and (7) are additionally taken into account, the final desired section is as shown in Fig. 4b. The transient *x*(*t*) in fuzzy ACS where FR is typed has a synthesized nonlinear characteristic as shown in Fig. 4c.

Based on the obtained FR nonlinear transformation and the above recommendations for the structure of the knowledge base, the appropriate content can be easily constructed:

- since the synthesized nonlinear transformation has four points of gain factor variation, the term set T_E for the linguistic error variable *E* will contain five fuzzy sets $\mathrm{T}_E = \{A_0, A_1, A_2, A_3, A_4\}$;
- the MFs specifying fuzzy sets A_0 and A_4 belong to classes *L* and γ respectively (due to the output constraint);
- the remaining MFs belong to class *t*.

As explained above, the definition area of the *i*th piecewise linear section of the regulator static characteristic depends on the mutual position of the adjacent MFs. The range depends on the value of the rule base conclusions from the rule base. Running through the entire definition area of the nonlinear characteristic (Fig. 4b) and taking into account its symmetry with respect to the origin of the coordinates, the MF parameters (Fig. 5) can be easily determined



(a) (b) (c)

Fig. 4. Yakubovich circular criterion (a), sections of absolute stability with nonlinear characteristic (b), and transient in ACS (c)

for the input variable, as well as the MF values for the output variable embedded in the rule base:

- IF E is A_0, THEN $u = 1.1$;
- IF E is A_1, THEN $u = 0.35$;
- IF E is A_2, THEN $u = 0$;
- IF E is A_3, THEN $u = -0.35$;
- IF E is A_4, THEN $u = -1.1$.

CONCLUSIONS

The present work develops the concept proposed by Makarov, according to which the fuzzy Zadeh transformation implemented in the circuit of type-1 FR ACS is in fact a nonlinear transformation. However, it becomes piecewise linear when the Sugeno model is used. For such a fuzzy system that uses methods from the theory of nonlinear control systems, the problem of dynamics research is solved in a form suitable for an engineer-developer. A methodology is proposed, which not only provides a definition of the ensured stability area, but can also be used to provide the required quality indices of the control process.

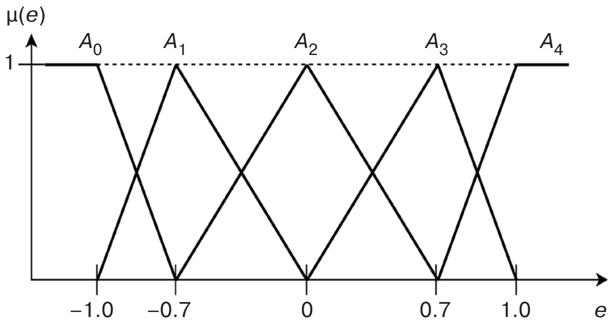


Fig. 5. Arrangement of the MF functions

Obviously, the results will also be valid when Mamdani, Larsen, and Tsukamoto fuzzy inference models are used in control systems: since the nonlinear transformations corresponding to these models are smooth, they can be approximated by piecewise linear sections. Thus, the solution of the problem can be reduced to the proposed method. In this case, the linear approximation section at the origin (having a small slope) is chosen based on the required steady-state error. Meanwhile, the slope of the steep section is selected based on the quality requirements according to the Yakubovich criterion. This approach can be used to solve analysis and synthesis problems of fuzzy ACS having a type-1 regulator.

Authors' contribution. All authors equally contributed to the research work.

REFERENCES

1. Makarov I.M., Lokhin V.M. *Intellektual'nye sistemy avtomaticheskogo upravleniya* (*Intelligent Automatic Control Systems*). Moscow: Fizmatlit; 2001. 576 p. (in Russ.). ISBN 978-5-9221-0162-2
2. Pospelov D.A. (Ed.). *Nechetkie mnozhestva v modelyakh upravleniya i iskusstvennogo intellekta* (*Fuzzy Sets in Control Models and Artificial Intelligence*). Moscow: Nauka; 1986. 312 p. (in Russ.).
3. Makarov I.M., Lokhin V.M., Manko S.V., Romanov M.P. *Iskusstvennyi intellekt i intellektual'nye sistemy upravleniya* (*Artificial Intelligence and Intelligent Control Systems*). Moscow: Nauka; 2006. 333 p. (in Russ.).
4. Makarov I.M., Lokhin V.M. *Artificial Intelligence and Complex Objects Control*. Lewiston: Edwin Mellen Press; 2000. 404 p.
5. Piegat A. *Fuzzy Modeling and Control*. Berlin: Physica Heidelberg; 2001. 728 p.
6. Makarov I.M., Lokhin V.M., Manko S.V., Romanov M.P., Sitnikov M.S. Stability of intellectual automatic control systems. *Informatsionnye tekhnologii = Information Technologies*. 2013;2:1–32 (in Russ.).
7. Rutkowska D., Pilinski M., Rutkowski L. *Neironnye seti, geneticheskie algoritmy i nechetkie sistemy* (*Neural Networks, Genetic Algorithms and Fuzzy Systems*): transl. from Pol. Moscow: Goryachaya liniya–Telekom; 2006. 452 p. (in Russ.). [Rutkowska D., Piliński M., Rutkowski L. *Sieci Neuronowe, Algorytmy Genetyczne i Systemy Rozmyte*. Warszawa; Łodż: Wydawnictwo Naukowe PWN. 2004.]
8. Hashemi S.M., Botez R. Lyapunov-based Robust Adaptive Configuration of the UAS-S4 Flight Dynamics Fuzzy Controller. *The Aeronautical Journal*. 2022;126(1301):1187–1209. https://doi.org/10.1017/aer.2022.2
9. Gandhi R., Adhyaru D. Takagi-Sugeno fuzzy regulator design for nonlinear and unstable systems using negative absolute eigenvalue approach. *IEEE/CAA Journal of Automatica Sinica*. 2020;7(2):482–493. https://doi.org/10.1109/JAS.2019.1911444
10. Lan L., Tiem N., Co Nhu V. Absolute Stability for a Class of Takagi-Sugeno Fuzzy Control Systems. In: *3rd International Conference on Robotics, Control and Automation Engineering* (*RCAE*). 2020. P. 47–51. https://doi.org/10.1109/RCAE51546.2020.9294352
11. Sakly A., Zahra B., Benrejeb M. Stability Study of Mamdani's Fuzzy Controllers Applied to Linear Plants. *Studies in Informatics and Control*. 2008;17(4):441–452.
12. Siddikov I., Porubay O., Rakhimov T. Synthesis of the neuro-fuzzy regulator with genetic algorithm. *Int. J. Electric. Comput. Eng.* (*IJECE*). 2024;14(1):184–191. http://doi.org/10.11591/ijece.v14i1.pp184-191
13. Hamza M., Yap I., Choudhury I. Genetic Algorithm and Particle Swarm Optimization Based Cascade Interval Type 2 Fuzzy PD Controller for Rotary Inverted Pendulum System. *Math. Probl. Eng.* 2015;2015(6). https://doi.org/10.1155/2015/695965
14. Mahmoodabadi M., Babak N. Robust fuzzy linear quadratic regulator control optimized by multi-objective high exploration particle swarm optimization for a 4 degree-of-freedom quadrotor. *Aerosp. Sci. Technol.* 2019;97:105598. https://doi.org/10.1016/j.ast.2019.105598
15. Sakalli A., Beke A., Kumbasar T. Gradient Descent and Extended Kalman Filter based self-tuning Interval Type-2 Fuzzy PID controllers. In: *2016 IEEE International Conference on Fuzzy Systems* (*FUZZ-IEEE*). 2016. P. 1592–1598. https://doi.org/10.1109/FUZZ-IEEE.2016.7737880
16. Islam S.U., Zeb K., Kim S. Design of Robust Fuzzy Logic Controller Based on Gradient Descent Algorithm with Parallel-Resonance Type Fault Current Limiter for Grid-Tied PV System. *Sustainability.* 2022;14(19):12251. https://doi.org/10.3390/su141912251
17. Bykovtsev Y.A. Synthesis of a Fuzzy Controller According to the Degree of Stability of the Control System. *Mekhatronika, Avtomatizatsiya, Upravlenie*. 2022;23(6):295–301 (in Russ.). https://doi.org/10.17587/mau.23.295-301
18. Bykovtsev Y.A., Lokhin V.M. Estimation of the accuracy of a control system with a fuzzy PID controller based on the approximation of the static characteristic of the controller. *Mekhatronika, Avtomatizatsiya, Upravlenie*. 2021;22(12):619–624. https://doi.org/10.17587/mau.22.619-624

СПИСОК ЛИТЕРАТУРЫ

1. Макаров И.М., Лохин В.М. *Интеллектуальные системы автоматического управления*. М.: Физматлит; 2001. 576 с. ISBN 978-5-9221-0162-2
2. Поспелов Д.А. (ред.). *Нечеткие множества в моделях управления и искусственного интеллекта*. М.: Наука; 1986. 312 с.
3. Макаров И.М., Лохин В.М., Манько С.В., Романов М.П. *Искусственный интеллект и интеллектуальные системы управления*. М.: Наука; 2006. 333 с.
4. Makarov I.M., Lokhin V.M. *Artificial Intelligence and Complex Objects Control*. Lewiston: Edwin Mellen Press; 2000. 404 p.
5. Piegat A. *Fuzzy Modeling and Control*. Berlin: Physica Heidelberg; 2001. 728 p.

6. Макаров И.М., Лохин В.М., Манько С.В., Романов М.П., Ситников М.С. Устойчивость интеллектуальных систем автоматического управления. *Информационные технологии*. 2013;2:1–32.
7. Рутковская Д., Пилиньский М., Рутковский Л. *Нейронные сети, генетические алгоритмы и нечеткие системы*: пер. с пол. М.: Горячая линия–Телеком; 2006. 452 с.
8. Hashemi S.M., Botez R. Lyapunov-based Robust Adaptive Configuration of the UAS-S4 Flight Dynamics Fuzzy Controller. *The Aeronautical Journal*. 2022;126(1301):1187–1209. https://doi.org/10.1017/aer.2022.2
9. Gandhi R., Adhyaru D. Takagi-Sugeno fuzzy regulator design for nonlinear and unstable systems using negative absolute eigenvalue approach. *IEEE/CAA Journal of Automatica Sinica*. 2020,7(2):482–493. https://doi.org/10.1109/JAS.2019.1911444
10. Lan L., Tiem N., Co Nhu V. Absolute Stability for a Class of Takagi-Sugeno Fuzzy Control Systems. In: *3rd International Conference on Robotics, Control and Automation Engineering* (*RCAE*). 2020. P. 47–51. https://doi.org/10.1109/RCAE51546.2020.9294352
11. Sakly A., Zahra B., Benrejeb M. Stability Study of Mamdani's Fuzzy Controllers Applied to Linear Plants. *Studies in Informatics and Control*. 2008;17(4):441–452.
12. Siddikov I., Porubay O., Rakhimov T. Synthesis of the neuro-fuzzy regulator with genetic algorithm. *Int. J. Electric. Comput. Eng.* (*IJECE*). 2024;14(1):184–191. http://doi.org/10.11591/ijece.v14i1.pp184-191
13. Hamza M., Yap I., Choudhury I. Genetic Algorithm and Particle Swarm Optimization Based Cascade Interval Type 2 Fuzzy PD Controller for Rotary Inverted Pendulum System. *Math. Probl. Eng.* 2015;2015(6). https://doi.org/10.1155/2015/695965
14. Mahmoodabadi M., Babak N. Robust fuzzy linear quadratic regulator control optimized by multi-objective high exploration particle swarm optimization for a 4 degree-of-freedom quadrotor. *Aerosp. Sci. Technol.* 2019;97:105598. https://doi.org/10.1016/j.ast.2019.105598
15. Sakalli A., Beke A., Kumbasar T. Gradient Descent and Extended Kalman Filter based self-tuning Interval Type-2 Fuzzy PID controllers. In: *2016 IEEE International Conference on Fuzzy Systems* (*FUZZ-IEEE*). 2016. P. 1592–1598. https://doi.org/10.1109/FUZZ-IEEE.2016.7737880
16. Islam S.U., Zeb K., Kim S. Design of Robust Fuzzy Logic Controller Based on Gradient Descent Algorithm with Parallel-Resonance Type Fault Current Limiter for Grid-Tied PV System. *Sustainability.* 2022;14(19):12251. https://doi.org/10.3390/su141912251
17. Быковцев Ю.А. Синтез нечеткого регулятора на основе оценки степени устойчивости системы управления. *Мехатроника, автоматизация, управление*. 2022;23(6):295–301. https://doi.org/10.17587/mau.23.295-301
18. Bykovtsev Y.A., Lokhin V.M. Estimation of the accuracy of a control system with a fuzzy PID controller based on the approximation of the static characteristic of the controller. *Mekhatronika, Avtomatizatsiya, Upravlenie*. 2021;22(12):619–624. https://doi.org/10.17587/mau.22.619-624

About the Authors

Yuri A. Bykovtsev, Cand. Sci. (Eng.), Assistant Professor, Department of Management Problems, Institute of Artificial Intelligence, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: bykovcev@mirea.ru. Scopus Author ID 57302607300, ResearcherID KRQ-5339-2024, RSCI SPIN-code 9961-4437, https://orcid.org/0009-0003-6671-5674

Valery M. Lokhin, Dr. Sci. (Eng.), Professor, Department of Management Problems, Institute of Artificial Intelligence, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). Laureate of the State Prize of the Russian Federation in Science and Technology. Laureate of the State Prize of the Russian Federation in Education. Member of the Scientific Council on Robotics and Mechatronics of the Russian Academy of Sciences. Honored Worker of Science of the Russian Federation. E-mail: kpu-mirea@yandex.ru. Scopus Author ID 6602931640, https://orcid.org/0000-0001-6708-9124

Об авторах

Быковцев Юрий Алексеевич, к.т.н., доцент, кафедра проблем управления, Институт искусственного интеллекта, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д 78). E-mail: bykovcev@mirea.ru. Scopus Author ID 57302607300, ResearcherID KRQ-5339-2024, SPIN-код РИНЦ 9961-4437, https://orcid.org/0009-0003-6671-5674

Лохин Валерий Михайлович, д.т.н., профессор, кафедра проблем управления, Институт искусственного интеллекта, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). Лауреат государственной премии РФ в области науки и техники. Лауреат премии Правительства РФ в области образования. Член научного Совета РАН по робототехнике и мехатронике. Заслуженный деятель науки РФ. E-mail: kpu-mirea@yandex.ru. Scopus Author ID 6602931640, https://orcid.org/0000-0001-6708-9124

Translated from Russian into English by K. Nazarov
Edited for English language and spelling by Thomas A. Beavitt

Multiple robots (robotic centers) and systems. Remote sensing and non-destructive testing

Роботизированные комплексы и системы. Технологии дистанционного зондирования неразрушающего контроля

UDC 623.09
https://doi.org/10.32362/2500-316X-2025-13-3-63-72
EDN SONRQG

RESEARCH ARTICLE

Identification system for detecting and suppressing unmanned aerial vehicles based on video interception for use in combat conditions

Alexey A. Konik [1],
Roman E. Afonin [1],
Ivan M. Klimov [2],
Dmitry S. Zinchenko [2, @]

[1] *I.D. Putilin Belgorod Law Institute of the Ministry of Internal Affairs of Russia, Belgorod, 308024 Russia*
[2] *FKU NGO "STiS", Ministry of Internal Affairs of Russia, Moscow, 111024 Russia*
[@] *Corresponding author, e-mail: dmitriy.zinchenko.1998@mail.ru*

• **Submitted:** 03.02.2025 • **Revised:** 17.02.2025 • **Accepted:** 21.03.2025

Abstract

Objectives. In the context of military operations, states face the threats of attacks by unmanned aerial vehicles (UAVs) on vulnerable assets, in particular, those pertaining to law enforcement agencies. Currently, there is no uniform or—more importantly—effective approach to detecting and suppressing certain types of UAVs, in particular, first person view (FPV) drones. The aim of the work is to develop modernized systems for the detection and suppression of enemy unmanned aerial vehicles and to justify their full-scale implementation in the service activities of law enforcement agencies.

Methods. The work used system-structural, comparative-legal, and measurement research methods along with analysis, observation, and field modeling. In addition, the research refers to generalized and systematized experience of using UAVs in combat conditions.

Results. The structure, basic tactical and technical characteristics are described according to the principle of operation of the identification complex for detecting and suppressing UAVs, having the ability to intercept analog radio signals carrying video information, which makes it possible to effectively detect and counteract enemy UAVs at a considerable distance. An algorithm of actions to be taken by law enforcement officers when using this complex is also developed and described. Proposals for creating a hardware and software system based on the complex with the possibility of spoofing a video stream are outlined.

Conclusions. The results of the study indicate the need to supply law enforcement agencies with an identification system for detecting and suppressing UAVs having the ability to intercept an analog radio signal carrying video information. The use of such a system across a wide area subject to UAV attacks will significantly improve the effectiveness of alerting all categories of employees and civilians, as well as contributing to the establishment of airspace control and improving the effectiveness of the fight against UAVs.

Keywords: identification system, analog radio signal, video information, unmanned aerial vehicle, antenna, descrambling, video stream

© A.A. Konik, R.E. Afonin, I.M. Klimov, D.S. Zinchenko, 2025

For citation: Konik A.A., Afonin R.E., Klimov I.M., Zinchenko D.S. Identification system for detecting and suppressing unmanned aerial vehicles based on video interception for use in combat conditions. *Russian Technological Journal.* 2025;13(3):63–72. https://doi.org/10.32362/2500-316X-2025-13-3-63-72, https://www.elibrary.ru/SONRQG

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Идентификационный комплекс обнаружения и подавления беспилотных воздушных судов на основе видеоперехвата для использования в условиях боевых действий

А.А. Коник [1],
Р.Е. Афонин [1],
И.М. Климов [2],
Д.С. Зинченко [2, @]

[1] *Белгородский юридический институт МВД России имени И.Д. Путилина, Белгород, 308024 Россия*
[2] *ФКУ НПО «СТиС» МВД России, Москва, 111024 Россия*
[@] *Автор для переписки, e-mail: dmitriy.zinchenko.1998@mail.ru*

• **Поступила:** 03.02.2025 • **Доработана:** 17.02.2025 • **Принята к опубликованию:** 21.03.2025

Резюме

Цели. В условиях проведения боевых действий государства столкнулись с угрозами атак беспилотных воздушных судов (БВС) на различные объекты, в частности, объекты силовых структур. В настоящее время нет единого, а самое главное, эффективного подхода по обнаружению и подавлению различных видов БВС, а именно, FPV-дронов. Целью работы является создание модернизированных комплексов обнаружения и подавления БВС противника и обоснование полномасштабного внедрения этих комплексов в служебную деятельность силовых структур.

Методы. В работе использовались системно-структурный, сравнительно-правовой, измерительный методы исследования, анализ, наблюдение и натурное моделирование, а также обобщение и систематизация опыта применения БВС в условиях боевых действий.

Результаты. Описаны структура, основные тактико-технические характеристики и принцип работы идентификационного комплекса обнаружения и подавления БВС с возможностью перехвата аналогового радиосигнала, несущего видеоинформацию, позволяющего эффективно обнаруживать и противодействовать БВС противника на значительном расстоянии. Разработан и описан алгоритм действий сотрудников силовых структур при работе с данным комплексом, изложены предложения по созданию на базе комплекса аппаратно-программных систем с возможностью подмены видеопотока.

Выводы. Полученные результаты исследования указывают на необходимость внедрения в служебную деятельность силовых структур идентификационного комплекса обнаружения и подавления БВС с возможностью перехвата аналогового радиосигнала, несущего видеоинформацию, применение которого на широком участке местности, подверженном атакам БВС, позволит существенно улучшить эффективность оповещения всех категорий служащих и гражданского населения, а также будет способствовать установлению контроля над воздушным пространством и повышению эффективности борьбы с БВС.

Ключевые слова: идентификационный комплекс; аналоговый радиосигнал, видеоинформация, беспилотное воздушное судно, антенна, дескремблирование, видеопоток

Для цитирования: Коник А.А., Афонин Р.Е., Климов И.М., Зинченко Д.С. Идентификационный комплекс обнаружения и подавления беспилотных воздушных судов на основе видеоперехвата для использования в условиях боевых действий. *Russian Technological Journal*. 2025;13(3):63–72. https://doi.org/10.32362/2500-316X-2025-13-3-63-72, https://www.elibrary.ru/SONRQG

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

The capabilities of modern unmanned aerial vehicles (UAVs) have been demonstrated in combat operations. UAV attacks carried out by copter-type UAVs equipped with a “drop” system and first person view (FPV) drone-type UAVs [1] pose a threat not only to civilian and industrial buildings and structures, but also to critical infrastructure facilities and facilities belonging to law enforcement agencies, as well as their personnel [2].

During military operations, the territories of many states are often subjected to indiscriminate and sometimes chaotic attacks by UAVs equipped with various munitions and explosive devices (PG-7V[1] (VL[2], VR[3], VM[4]), RKG-3[5], etc.) [3], especially improvised explosive devices (Fig. 1).

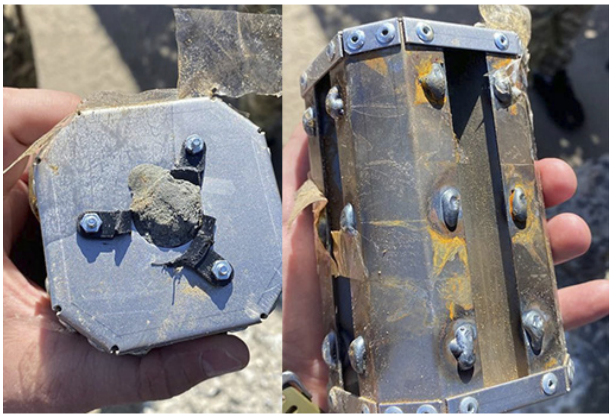


Fig. 1. Improvised explosive devices used in an UAV attack

In addition, recently developed “drop” systems for FPV drones are already in widespread use. These are capable of carrying out multiple attacks using different types of modernized munitions [4]: for example, FPV drones can remotely mine terrain by dropping munitions of different types: “butterfly,” “bell,” etc.

The above factors, together with the analysis of statistical data, indicate that FPV drones are the most dangerous to the civil population, law enforcement personnel, and facilities. They can attack at high speeds (over 100km/h), carry a significant payload in the form of an explosive charge, and penetrate deep into territory to a distance of 15–20 km, and in certain cases, up to 30 km [5]. Moreover, combat experience has shown that contemporary FPV drones use non-standard, permanently shifted frequencies for control [6], significantly increasing their resistance to electronic reconnaissance and electronic warfare (EW) detection.

In view of the increasingly challenging situation and the intensified use of FPV drones, it is necessary to introduce additional measures to increase the effectiveness of countering the enemy in this context, as well as to inform the civilian population of the threat of UAV attacks in a timely manner. The actual direction is the development of improved complexes for detection and suppression of UAVs. In particular, the creation and introduction of an identification complex for detection and suppression of UAVs with the ability to intercept analog radio signals carrying video information [7].

IDENTIFICATION COMPLEX FOR UAV DETECTION AND SUPPRESSION, CAPABLE OF INTERCEPTING ANALOG RADIO SIGNALS CARRYING VIDEO INFORMATION

Based on the above considerations, as well as electronic reconnaissance combat operations data confirming it to be a key EW approach and effective circuit of destroying enemy UAVs [8], researchers from the Belgorod oblast of the Russian Federation have developed an identification complex for detecting and suppressing UAV (hereinafter “the complex”). The complex can be used to detect and identify an enemy UAV in the airspace with a high degree of probability. The complex is based on the interception of the analog radio signal carrying video information from enemy UAVs.

[1] A shot with an anti-tank grenade.
[2] A shot with an anti-tank grenade “Luch.”
[3] A shot with an anti-tank grenade “Resume.”
[4] A modernized anti-tank grenade shot.
[5] A handheld shaped-charge grenade.

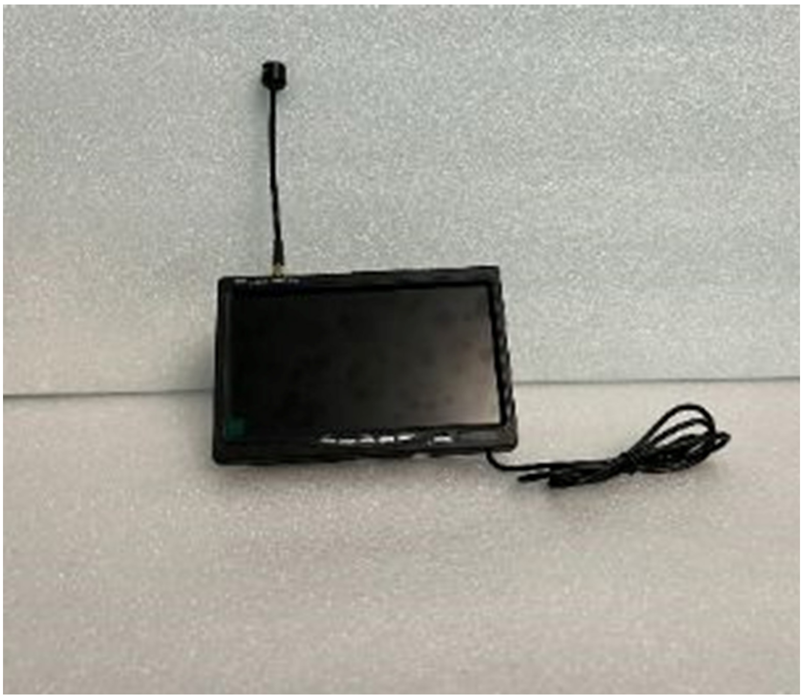


Fig. 2. "EFIR" UAV detector

It should be noted that similar products are currently being developed for this purpose, such as the "EFIR" and "Umbrella" detectors (Fig. 2). The basic operating principle of such complexes is based on scanning the air and recording the carrier frequency with detecting and recording the video stream.

However, from an analysis and evaluation of the tactical and technical characteristics of such products, it appears that they do not allow the simultaneous detection of UAV signals in a wide range of frequencies (the products include a minimum number of frequency bands, for example 2.4 and 5.8 GHz) and at long distances (the declared maximum detection distance ranges from 500 m to 5 km) [9].

The developed identification complex includes the following components:

- directional receiving antennas for specific frequency ranges (circular antennas can be used, but in this case the detection distance is much shorter);
- receiver for analog radio signal carrying video information;
- video signal transmission cable with RCA connectors (the length of the cable depends on the installation configuration of the equipment, and it should be noted that the quality of data transmission decreases with increasing cable length);
- monitor with VGA or HDMI connector;
- AV RCA-VGA or AV RCA-HDMI monitor connection converter (depending on the monitor used);
- VGA-VGA or HDMI-HDMI cable;
- 2 USB, 5V/2.1A power supplies.

The complex operating principle consists in the use of a circular antenna (when the direction of the likely UAV take-off is unknown) or a directional antenna to continuously monitor the surrounding space within a radius of 1.5 to 3 km (when using a circular antenna) and up to 30 km (when using a directional antenna). These distances depend on a number of factors such as the transmitting power of the FPV drone, the gain of the receiving antenna, the terrain, and so on. When the UAV with analog video signal transmission channel appears, the complex automatically intercepts and transmits (duplicates) to the monitor a video with an illustration of the terrain over which the UAV is flying and flight data, which is observed on the monitor screen or in the FPV drone goggles by the operator who is controlling it.

The developed complex is capable of operating across a wide range of temperatures, under strong vibrations, and in various other extreme conditions. An important aspect of the complex's operation is its passivity, meaning that it does not emit any signals of its own that could be detected by devices that scan radio waves.

A special feature of the complex is its ability not only to detect UAVs by receiving an analog video signal (displayed on the monitor), but also to descramble encrypted "X-signals." This allows for the prior identification of illegal flights of UAVs, with the possibility of determining their speed and altitude, predicting their trajectories, and identifying possible targets. The detection range of enemy UAVs (up to 30 km) allows time for security and countermeasures to be put into place.

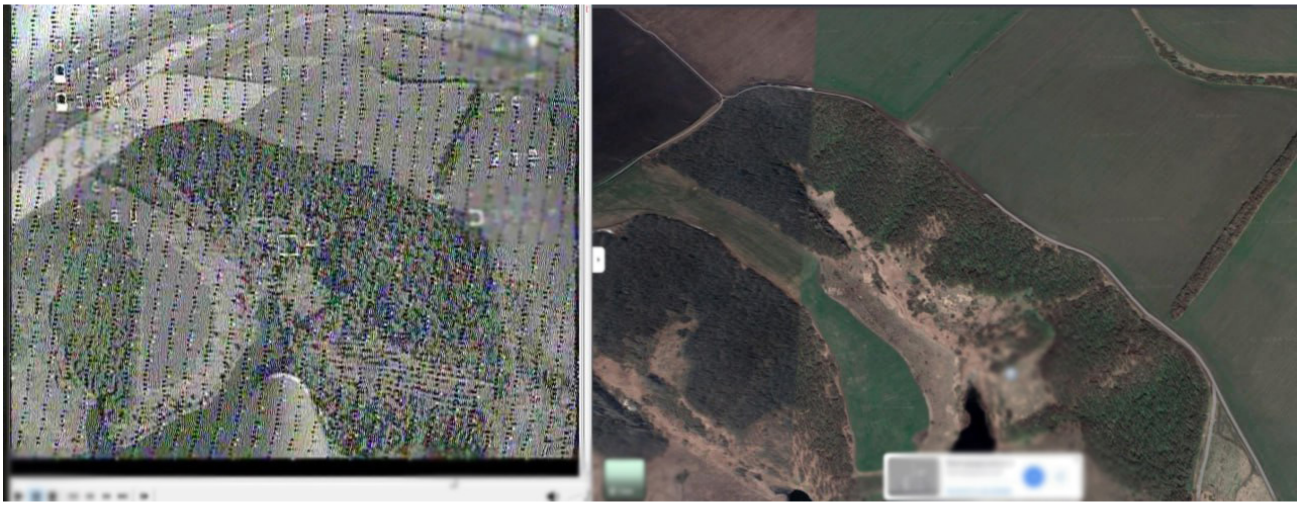


Fig. 3. Identifying UAV flight directions by collating complex data on terrain maps

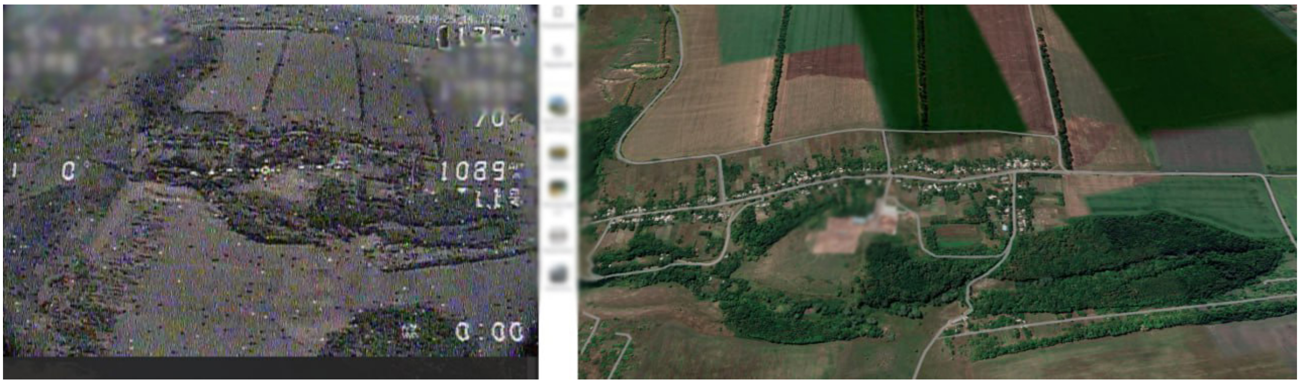


Fig. 4. Capturing UAV analog video and linking the flight path to a terrain map

It should also be noted that in addition to the above characteristics, the complex offers unique capabilities and potential for use in combat operations. However, for reasons of confidentiality, the circuitry and more detailed tactical and technical characteristics of the complex cannot be disclosed.

PRACTICAL APPLICATION OF THE DEVELOPED IDENTIFICATION COMPLEX FOR UAV DETECTION AND SUPPRESSION

After deploying the developed complex at a site, its successfully tactical and technical characteristics were confirmed by detecting more than 500 UAVs over the course of several months of operation. For example, in just one day, the complex enabled the detection of 20 FPV drones, 9 of which were successfully suppressed[6].

The complex was used to identify UAV flight directions by analyzing topographic maps, including offline electronic maps (Figs. 3 and 4). In addition, under certain conditions, it is possible to determine UAV launch points. Furthermore, the integration of the complex into the mobile version is being actively pursued in order to use the complex on mobile objects.

In addition to military applications, the complex has been successfully used by law enforcement officers. Over a short period of time, it has been used to significantly increase the efficiency of detection and suppression of UAVs in the controlled area (18 FPV drones destroyed). The practical use of the complex has demonstrated that an integral part of its successful application is the maintenance of appropriate documentation (log, statement, etc.), which must necessarily reflect the information received (location of the UAV detected with reference to a settlement or exact coordinates, the nature of the UAV behavior, the measures taken, and the result obtained, etc.). To this end, researchers from the FKU NPO "STiS"[7] of the Ministry of Internal Affairs of Russia and the I.D. Putilin Belgorod Law Institute of the Ministry of Internal Affairs of the Russian Federation[8] have developed an optimal template for the logbook of data on the detection of illegal UAVs.

[6] Statistical data (materials) are provided by the initiative group, the developers of the complex.

[7] https://стис.мвд.рф (in Russ.). Accessed March 20, 2025.

[8] https://белюи.мвд.рф (in Russ.). Accessed March 20, 2025.

Therefore, for the timely detection and effective suppression of hostile UAVs, it is proposed to introduce this complex into the official activities of law enforcement officers and create a network of strongholds equipped with the complex alongside EW and other means of suppression of UAVs on the basis of established roadblocks, checkpoints, and dislocations, as well as at temporary bases of law enforcement agencies with further alerting of the authorities through closed communication channels and the civilian population through open ones (e.g., in messenger groups etc.).

It is also necessary to define the algorithm of actions for law enforcement officers [10] when using the complex:

- identifying the threat;
- alerting personnel, operations duty officer, countermeasures group, and civilians through prepared communication channels;
- taking measures to suppress the enemy UAV by all possible means (EW, physical elimination including the use of firearms [11], etc.);
- taking measures to eliminate the consequences of UAV suppression or detonation and documenting the incident.

Personnel should be alerted simultaneously with civilians. The message should be brief but informative: UAV direction, expected type, required action.

UAVs can be suppressed either directly in the vicinity of the observation station, or using the tactics of mobile ambush groups armed with additional electronic reconnaissance equipment and various UAV suppression equipment (EW systems, smoothbore weapons, etc.) to intercept and suppress drones away from civilian areas and critical assets.

It should be borne in mind that the use of EW means in the immediate vicinity of the complex may be the cause of interference and equipment malfunction. Therefore, it is necessary to switch off the various EW complexes and other means of UAV suppression in the immediate vicinity of the complex during its use. Since EW cannot guarantee UAV suppression [12], it is also necessary to be prepared to use firearms to defeat the drone. At low altitudes (up to 40 m), UAVs can be effectively suppressed with smoothbore weapons loaded with shotgun cartridges #3 and #5. Alternatively, an effective "density" of fire can be created with 3–5 automatic weapons [13].

Once the UAV has been disposed of, first aid should be administered to any casualties and the scene cordoned off from unauthorized persons, taking care to ensure that no further mined parts remain on the surviving parts of the UAV. Taking all necessary safety precautions, the surviving parts of the UAV should then be reassembled for examination. If the status of the UAV is in doubt, or if it has been destroyed without an explosive device detonating, it is strictly forbidden to approach it until the object has been examined by experts. This is because the explosive device may be remotely triggered or rigged to detonate after a certain period of time.

In addition, regular briefings are required to develop the knowledge and skills of the personnel in the algorithm of actions to be taken upon detection of a UAV.

STRATEGIC TASKS SOLVED BY THE DEVELOPED IDENTIFICATION COMPLEX FOR DETECTING AND SUPPRESSING UAVS

One of the strategic goals of using the complex consists in the collection of information, followed by its mandatory analysis and use in the official activities of law enforcement officers, including special forces.

When used on a regular basis, the system also solves the following tasks:

- UAV direction detection and tracking. This allows UAV flight paths to be tracked and, in some cases, the launching point and the most dangerous directions to be determined and recorded in surveillance logs and terrain maps, including the use of various services. The data thus obtained can be used to reinforce the EW systems in certain directions and to carry out ambush operations;
- identification of UAV models and types [14]. The nature of the drones behavior in the air (flight speed and altitude, stability, etc.) allows UAV models and types to be identified with a high degree of accuracy, making it possible to predict further operator actions and prepare for a potential threat;
- detection of enemy groups (with UAV identity or unit call sign information) conducting terrorist activities in the vicinity (in a specific area). This information is often displayed on screen when an analog radio signal carrying video information is intercepted. After the attack, this data could potentially be passed on to the investigating authorities for analysis. In addition, this approach allows a more accurate determination of the number and size of enemy UAV groups in a given area;
- identification of criminals (traitors). At present, there are known cases of launching UAVs by hostile sabotage-reconnaissance groups and criminals (traitors) from various territories. Since these facts are difficult to detect without special equipment, such groups may operate for a long time, causing selective damage to vital infrastructure and civilian population. With a high-quality functional arrangement of the complex, the UAV take-off position can be captured by a video signal (when entering the antenna coverage area), thus enabling

measures to be taken to liquidate and eliminate such groups;

- documentation of UAV suppression (elimination) or the result of using it. In this case, it is a matter of documenting both the disposal of the UAV and the further collection of its fragments, including in the case of its suppression without disposal, for further investigation. There are also known cases of drones being found by civilians and injured by a detonation of the explosive device. In the event of UAV attacks and injuries, the complex enables rapid response and dispatch of the necessary services to the designated location to provide assistance;
- determination of UAV technical abilities [15]. Using the complex with the ability to intercept the analog radio signal carrying video information, it is possible to determine the flight range and limit distances. It is also possible to monitor new techniques used by the enemy (e.g., the use of special software makes it possible to record the frequency values of the UAV used during the flight), including any increase in battery power. The complex can also be used to determine the resistance of UAVs to the EW systems used to counter them and make adjustments to its operation.

CONCLUSIONS

The use of an identification complex for the detection and suppression of UAVs having the ability to intercept analog radio signals carrying video information, along with the possibility to protect a large area of terrain vulnerable to UAV attack, significantly improves the effectiveness of alerting all categories of officials and the civilian population. It also contributes to the establishment of airspace control to increase the effectiveness of the fight against UAVs. In turn, the collection and analysis of data obtained using the complex and recorded in the logbook can be used to supply accurate forecasts of the situation in a given area.

On the basis of the characteristics identified by scientific analysis and practical application of the developed complex, the following innovative features that determine its potential in conditions of special military operation can be distinguished:

- wide frequency range (more than 4 frequencies);
- UAV detection range up to 30 km;
- the ability to descramble encrypted “X” signals;
- resistance to extreme conditions and a wide operating temperature range: from −25 to +40°C.

Thus, the results of the study determine the possibility of creating hardware-software systems with the possibility of spoofing the video stream, which implies interference in the form of transmitting an extraneous video to the UAV operator, as well as permitting various modifications of the complex: stationary, automotive, etc.

Authors’ contributions

A.A. Konik—creation of the research concept and methodology, preparation of the initial version of the manuscript, literature review, editing of the manuscript text.

R.E. Afonin—development of methodology, preparation of the initial version of the manuscript, collection and systematization of information on the practice of using the complex, literature review.

I.M. Klimov—analysis of existing solutions, providing recommendations on problem formulation and development of the complex, conducting an expert assessment of the results obtained.

D.S. Zinchenko—methodology development, preparation of the initial version of the manuscript, interpretation and generalization of the results, general guidance, preparation of the article for publication.

All the authors made a common equivalent contribution, read and agreed on the published version of the manuscript.

REFERENCES

1. Ruznyaev E.S. Foreign unmanned aerial vehicles for military purposes. *Matritsa Nauchnogo Poznaniya*. 2022;6(2):42–47 (in Russ.).
2. Pavlov R.A., Saveliev R.A. The use of unmanned aerial vehicles in modern military conflicts. *Molodoi uchenyi = Young Scientist*. 2022;51(446):48–50 (in Russ.).
3. Averchenko S.V., Belousov V.V. Unmanned aerial vehicles in military conflicts of the second half of the 20th – beginning of the 21st centuries: the main milestones of history. *Sovremennaya nauchnaya mysl' = Modern Scientific Thought*. 2023;1:231–242 (in Russ.).
4. Zinchenko D.S., Afonin R.E. Analysis of means and methods of passive protection against attacks by unmanned aircraft in the context of a special military operation. *Problemy pravookhranitel'noi deyatel'nosti = Problems of Law Enforcement*. 2024;4(58):56–64 (in Russ.). https://elibrary.ru/ibjkhy
5. Mataras A.A., Gulyaev I.Yu. Analysis of the use of FPV drones during combat operations 2014–2023. In: *Current Issues of Increasing Effective Fire Training in Law Enforcement Agencies: Theory and Practice* (*The Third Makarov Readings*): *All-Russian Collection of Scientific and Practical Materials.* Perm; 2023. V. 3. P. 135–141 (in Russ.). https://elibrary.ru/xcnjax
6. Nikolaev N.V., Ilyin V.V., Nekrasov M.I. Actual issues of countering modern autonomous unmanned aerial vehicles and FPV drones. *Voprosy bezopasnosti = Security Issues*. 2024;1:40–56 (in Russ.).
7. Suleimanov M.V., Konyaev V.M., Guts S.I. Methods of detecting small commercial UAVs. In: *Means and Systems for Countering Unmanned Aerial Vehicles: Collection of Reports of the Scientific and Practical Special Conference*. Moscow. 2023. P. 22–26 (in Russ.). https://www.elibrary.ru/fcxjpg
8. Krasnenko N.P., Bogushevich A.Ya., Kurakov S.A., Rakov A.S., Rybakov I.A. Detection of unmanned aerial vehicles: existing solutions and possibilities. In: *All-Russian Open Armand Readings: Modern Problems of Remote Sensing, Radar, Wave Propagation and diffraction*. 2024. P. 429–438 (in Russ.). https://doi.org/10.24412/2304-0297-2024-1-429-440
9. Falileev V.Yu., Shatovkin R.R. Analysis of existing automated protection against drones. *Vozdushno-kosmicheskie sily. Teoriya i praktika = Aerospace Forces. Theory and Practice*. 2020;14:130–140 (in Russ.). https://www.elibrary.ru/aypwcg
10. Konik A.A., Klimov I.M. On the issue of means and methods of active protection against attacks by unmanned aircraft in the context of a special military operation. *Problemy pravookhranitel'noi deyatel'nosti = Problems of Law-Enforcement Activity*. 2024;4(58):65–69 (in Russ.). https://www.elibrary.ru/kmijke
11. Makarenko S.I., Timoshenko A.V. Counter Unmanned Aerial Vehicles. Part 2. Rocket and Artillery Fire, Physical Interception. *Sistemy upravleniya, svyazi i bezopasnosti = Systems of Control, Communication and Security*. 2020;1:147–197 (in Russ.).
12. Tikshaev V.N., Barvinenko V.V. The problem of combating unmanned aerial vehicles and possible solutions. *Voennaya mysl' = Military Thought*. 2021;1:125–131 (in Russ.).
13. Moiseev N.V. Prospective variants of BVS lesions. In: *Modern Russian Science: Current Issues, Achievements and Innovations: collection of articles of the VI All-Russian Scientific and Practical Conference*. Penza: Nauka i Prosveshchenie; 2023. P. 106–109 (in Russ.).
14. Berdymuratov D.B. Analysis of UAV identification systems and database development. *UNIVERSUM: Technical Sciences*. 2024;10(127):4–8 (in Russ.).
15. Khrapkov D.S., Romashov V.A. Unmanned aerial vehicles as modern means of conducting combat operations. *Nauchnaya mysl' = Scientific thought*. 2021;15(1-1(39)):69–71 (in Russ.). https://www.elibrary.ru/ootziq

СПИСОК ЛИТЕРАТУРЫ

1. Рузняев Е.С. Зарубежные беспилотные летательные аппараты военного назначения. *Матрица научного познания*. 2022;6(2):42–47.
2. Павлов Р.А., Савельев К.П. Применение беспилотных летательных аппаратов в современных военных конфликтах. *Молодой ученый*. 2022;51(446):48–50.
3. Аверченко С.В., Белоусов В.В. Беспилотные летательные аппараты в военных конфликтах второй половины XX – начала XXI веков: основные вехи истории. *Современная научная мысль*. 2023;1:231–242.
4. Зинченко Д.С., Афонин Р.Е. Анализ средств и способов пассивной защиты от атак беспилотных воздушных судов в условиях проведения специальной военной операции. *Проблемы правоохранительной деятельности*. 2024;4(58):56–64. https://elibrary.ru/ibjkhy
5. Матарас А.А., Гуляев И.Ю. Анализ применения FPV дронов в ходе боевых действий 2014–2023 гг. В сб.: *Актуальные вопросы повышения эффективной огневой подготовки в силовых структурах: теория и практика* (*III Макаровские чтения*): *Всероссийский сборник научно-практических материалов*. Пермь. 2023. Т. 3. С. 135–141. https://elibrary.ru/xcnjax
6. Николаев Н.В., Ильин В.В., Некрасов М.И. Актуальные вопросы противодействия современным автономным беспилотным летательным аппаратам и FPV-дронам. *Вопросы безопасности*. 2024;1:40–56.
7. Сулейманов М.В., Коняев В.М., Гуц С.И. Способы обнаружения малых БВС коммерческого типа. В сб.: *Средства и системы противодействия беспилотным воздушным судам: Сборник докладов научно-практической специальной конференции*. Москва. 2023. С. 22–26. https://www.elibrary.ru/fcxjpg
8. Красненко Н.П., Богушевич А.Я., Кураков С.А., Раков А.С., Рыбаков И.А. Обнаружение беспилотных летательных аппаратов: существующие решения и возможности. В сб.: *Всероссийские открытые Армандовские чтения: Современные проблемы дистанционного зондирования, радиолокации, распространения и дифракции волн*. Муром. 2024. С. 429–438. https://doi.org/10.24412/2304-0297-2024-1-429-440

9. Фалилеев В.Ю., Шатовкин Р.Р. Анализ существующих автоматизированных комплексов защиты от дронов. *Воздушно-космические силы. Теория и практика*. 2020;14:130–140. https://www.elibrary.ru/aypwcg
10. Коник А.А., Климов И.М. К вопросу о средствах и способах активной защиты от атак беспилотных воздушных судов в условиях проведения специальной военной операции. *Проблемы правоохранительной деятельности*. 2024;4(58):65–69. https://www.elibrary.ru/kmijke
11. Макаренко С.И., Тимошенко А.В. Анализ средств и способов противодействия беспилотным летательным аппаратам. Часть 2. Огневое поражение и физический перехват. *Системы управления, связи и безопасности*. 2020;1:147–197.
12. Тикшаев В.Н., Барвиненко В.В. Проблема борьбы с беспилотными летательными аппаратами и возможные пути ее решения. *Военная мысль*. 2021;1:125–131.
13. Моисеев Н.В. Перспективные варианты поражения БВС. В сб.: *Современная Российская наука: актуальные вопросы, достижения и инновации: сборник статей VI Всероссийской научно-практической конференции*. Пенза: Наука и Просвещение; 2023. С. 106–109.
14. Бердымуратов Д.Б. Анализ систем идентификации и формирования базы данных беспилотных летательных аппаратов. *Universum: технические науки*. 2024;10(127):4–8.
15. Храпков Д.С., Ромашов В.А. Беспилотные летательные аппараты, как современные средства ведения боевых действий. *Научная мысль*. 2021;15(1-1(39));69–71. https://www.elibrary.ru/ootziq

About the Authors

Alexey A. Konik, Cand. Sci. (Ped.), Associate Professor, Deputy Head of the Department of Tactical and Special Training, I.D. Putilin Belgorod Law Institute of the Ministry of Internal Affairs of Russia (71, Gor'kogo ul., Belgorod, 308024 Russia). E-mail: 89205666067@mail.ru. RSCI SPIN-code 9074-3701, https://orcid.org/0000-0003-4563-3509

Roman E. Afonin, Senior Lecturer, Department of Physical Training, I.D. Putilin Belgorod Law Institute of the Ministry of Internal Affairs of Russia (71, Gor'kogo ul., Belgorod, 308024 Russia). E-mail: afonin.roman@mail.ru. RSCI SPIN-code 4015-0302, https://orcid.org/0009-0000-6383-3913

Ivan M. Klimov, Head of the Department for the Development of Means to Counter the Functioning of Unmanned Vehicles, Scientific and Production Association "Special Equipment and Communications," Ministry of Internal Affairs of the Russian Federation (2, Prud Klyuchiki ul., Moscow, 111024 Russia). E-mail: klimov.ivan700@gmail.com. https://orcid.org/0009-0006-9824-8224

Dmitry S. Zinchenko, Senior Researcher, Department for the Development of Means to Counter the Functioning of Unmanned Vehicles, Scientific and Production Association "Special Equipment and Communications," Ministry of Internal Affairs of the Russian Federation (2, Prud Klyuchiki ul., Moscow, 111024 Russia). E-mail: dmitriy.zinchenko.1998@mail.ru. RSCI SPIN-code 9816-4245, https://orcid.org/0009-0001-0948-8239

Об авторах

Коник Алексей Алексеевич, к.пед.н., доцент, заместитель начальника кафедры тактико-специальной подготовки, ФГКОУ ВО «Белгородский юридический институт Министерства внутренних дел Российской Федерации имени И.Д. Путилина» (308024, Россия, Белгород, ул. Горького, д. 71). E-mail: 89205666067@mail.ru. SPIN-код РИНЦ 9074-3701, https://orcid.org/0000-0003-4563-3509

Афонин Роман Евгеньевич, старший преподаватель, кафедра физической подготовки, ФГКОУ ВО «Белгородский юридический институт Министерства внутренних дел Российской Федерации имени И.Д. Путилина» (308024, Россия, Белгород, ул. Горького, д. 71). E-mail: afonin.roman@mail.ru. SPIN-код РИНЦ 4015-0302, https://orcid.org/0009-0000-6383-3913

Климов Иван Михайлович, начальник отдела развития средств противодействия функционированию беспилотных аппаратов, Центр развития беспилотных аппаратов и средств противодействия их функционированию, Научного-исследовательский институт специальной техники, ФКУ «Научно-производственное объединение «Специальная техника и связь», Министерство внутренних дел Российской Федерации» (ЦРБАиСПиФ НИИСТ ФКУ НПО «СТиС» МВД России) (111024, Россия, Москва, ул. Пруд Ключики, д. 2). E-mail: klimov.ivan700@gmail.com. https://orcid.org/0009-0006-9824-8224

Зинченко Дмитрий Сергеевич, старший научный сотрудник, отдел развития средств противодействия функционированию беспилотных аппаратов, Центр развития беспилотных аппаратов и средств противодействия их функционированию, Научного-исследовательский институт специальной техники, ФКУ «Научно-производственное объединение «Специальная техника и связь», Министерство внутренних дел Российской Федерации» (ЦРБАиСПиФ НИИСТ ФКУ НПО «СТиС» МВД России) (111024, Россия, Москва, ул. Пруд Ключики, д. 2). E-mail: dmitriy.zinchenko.1998@mail.ru. SPIN-код РИНЦ 9816-4245, https://orcid.org/0009-0001-0948-8239

Translated from Russian into English by K. Nazarov
Edited for English language and spelling by Thomas A. Beavitt

Modern radio engineering and telecommunication systems

Современные радиотехнические и телекоммуникационные системы

UDC 621.396.96
https://doi.org/10.32362/2500-316X-2025-13-3-73-83
EDN XDZDDH

RESEARCH ARTICLE

On the equivalence of characteristics and specularity in the construction of traditional and MIMO radars with a parallel view of space based on antenna arrays

Boris M. Vovshin [1, 2, @],
Alexander A. Pushkov [1],
Elizaveta M. Khalturina [1]

[1] NPO ALMAZ, Moscow, 127411 Russia
[2] MIREA – Russian Technological University, Moscow, 119454 Russia
[@] Corresponding author, e-mail: boris@eleron.net

• **Submitted:** 30.07.2024 • **Revised:** 13.02.2025 • **Accepted:** 31.03.2025

Abstract

Objectives. In recent years, more and more attention has been paid in radar theory and practice to the development of multiple-input and multiple-output (MIMO) radar, which offers a number of advantages over traditional radar based on phased antenna arrays (PAAs). These include the possibility to flexibly view space and adapt to a changing signal-interference environment, etc. MIMO technology used in radar requires the emission of a probe signal in the form of a coherent system of orthogonal signals, each of which triggers its own emitter in the transmitting antenna array (AA). As a result, the specified target search area is simultaneously illuminated. Specific spatiotemporal processing (SSP) is used to collect signals from all directions in the irradiated zone at the receiver output. In this regard, the task of finding an SSP structure in MIMO radar that is optimal compared to the traditional approach becomes urgent. The study set out to synthesize the structure of SSP with single–channel reception in MIMO radar and compare the obtained structure and characteristics with those similar in traditional parallel-view radars based on multipath receiving radar.

Methods. The study is based on methods and principles of the theory of multibeam synthesized aperture antennas and methods for the synthesis of optimal Neiman–Pearson detectors based on the likelihood ratio.

Results. For a MIMO radar with AA for transmission and reception provided by a single weakly directional antenna, a split SSP was synthesized to form optimal pre-threshold statistics (PTS) of the detector against a background of white Gaussian noise. The obtained PTS is compared with a similar PTS in a traditional parallel space survey radar with a mirror structure.

Conclusions. It is shown that the detection quality indicators of the compared radars in the mirror construction are equivalent in the mode of parallel target search in the same spatial sectors.

Keywords: MIMO radar, parallel space survey, space-time processing, FFT algorithm, multipath antenna array, pre-threshold statistics

© B.M. Vovshin, A.A. Pushkov, E.M. Khalturina, 2025

For citation: Vovshin B.M., Pushkov A.A., Khalturina E.M. On the equivalence of characteristics and specularity in the construction of traditional and MIMO radars with a parallel view of space based on antenna arrays. *Russian Technological Journal.* 2025;13(3):73–83. https://doi.org/10.32362/2500-316X-2025-13-3-73-83, https://www.elibrary.ru/XDZDDH

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Об эквивалентности характеристик и «зеркальности» построения традиционных и MIMO радиолокационных станций при параллельном обзоре пространства на основе антенных решеток

Б.М. Вовшин $^{1,\,2,\,@}$**,**
А.А. Пушков ^{1}**,**
Е.М. Халтурина ^{1}

[1] НПО «Алмаз», Москва, 127411 Россия
[2] МИРЭА – Российский технологический университет, Москва, 119454 Россия
[@] Автор для переписки, e-mail: boris@eleron.net

• **Поступила:** 30.07.2024 • **Доработана:** 13.02.2025 • **Принята к опубликованию:** 31.03.2025

Резюме

Цели. В последние годы в теории и практике радиолокации все больше внимания уделяется вопросам создания MIMO (англ., «много входов – много выходов») радиолокационных станций (РЛС), обладающих рядом достоинств перед традиционными РЛС с фазированными антенными решетками. К этим достоинствам следует отнести возможности гибкого обзора пространства, адаптации к меняющейся сигнально-помеховой обстановке и т.д. Технология MIMO в радиолокации требует излучения зондирующих сигналов в виде когерентной системы ортогональных сигналов, каждый из которых возбуждает собственный излучатель передающей антенной решетки (АР). Вследствие этого одновременно «освещается» заданная зона поиска цели. Пространственно-временная обработка (ПВО) «собирает» сигналы со всех направлений в облученной зоне на выходе приемника. В связи с этим актуальной является задача поиска оптимальной структуры ПВО в MIMO РЛС по сравнению с традиционным подходом. Цель работы – синтез структуры ПВО при одноканальном приеме в MIMO РЛС и сравнение полученного построения и характеристик с аналогичными в традиционных РЛС параллельного обзора на основе многолучевой приемной АР.

Методы. Использованы методы и принципы теории многолучевых антенн с синтезированной апертурой и методы синтеза оптимальных по критерию Неймана – Пирсона обнаружителей на основе отношения правдоподобия.

Результаты. Для MIMO РЛС с АР на передачу и одиночной слабонаправленной антенной на прием синтезирована разделяющаяся ПВО, формирующая оптимальную предпороговую статистику (ППС) обнаружителя на фоне белого гауссова шума. Проведено сравнение полученной ППС с аналогичной ППС в традиционной РЛС параллельного обзора пространства, имеющей «зеркальное» построение.

Выводы. Доказано, что в режиме параллельного поиска цели в одинаковых пространственных секторах показатели качества обнаружения у сравниваемых РЛС при «зеркальном» построении эквивалентны.

Ключевые слова: MIMO РЛС, параллельный обзор пространства, пространственно-временная обработка, алгоритм БПФ, многолучевая антенная решетка, предпороговая статистика

Для цитирования: Вовшин Б.М., Пушков А.А., Халтурина Е.М. Об эквивалентности характеристик и «зеркальности» построения традиционных и MIMO радиолокационных станций при параллельном обзоре пространства на основе антенных решеток. *Russian Technological Journal*. 2025;13(3):73–83. https://doi.org/10.32362/2500-316X-2025-13-3-73-83, https://www.elibrary.ru/XDZDDH

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

In recent years, research into Multiple-Input, Multiple-Output (MIMO) radars has claimed an important position within the theory and practice of radiolocation. Interest in MIMO radars has arisen in connection with the emerging possibilities of overcoming the limitations of traditional phased antenna array (PAA) radars in observing targets. MIMO technology is anticipated to be as revolutionary as the electronic scanning that replaced mechanical scanning in antenna technology to provide new radar characteristics and functionality [1, 2].

The idea of MIMO radar was originally based on the well-known property of radars in survey mode: the signal-to-noise ratio (SNR) at the receiver input and consequent detection quality index is practically independent of the transmit beamwidth $\Delta\theta_{0.5\text{tr}}$ for a given survey sector $\Delta\theta_{\text{svy}}$ and time t_{svy}. This statement, which is based on the fact that a decrease of the antenna directivity on transmission and concomitant decrease in the signal level on the target in the radar can be compensated by increasing the observation time, can be clarified as follows.

During the target location time in the beam width $\Delta\theta_{0.5}$, the size of the accumulated packet of reflected signals with period T_0 is $Q = (t_{\text{svy}}\Delta\theta_{0.5})/(T_0\Delta\theta_{\text{svy}})$. Therefore, as the observation rate increases, the value of Q decreases in proportion to the decrease in t_{svy}, which can only be increased at $T_0 = \text{const}$ by widening the beam by $\Delta\theta_{0.5}$. If the beam width is matched to the survey sector $\Delta\theta_{0.5} = \Delta\theta_{\text{svy}}$, the echo signals from all targets having a priori unknown angular coordinates within the coverage area can be collected using a receiving multibeam AA (MBAA). When this parallel type of view is applied, the number of coherently accumulated pulses Q is limited only by the correlation time interval of the target itself $t_{\text{TRG corr}}$ [3, 4]. In practice, traditional radars with parallel-view PAA have the following well-known disadvantages [5, 6]:

1. A wide directivity pattern (DP) for transmission equal to $\Delta\theta_{\text{svy}}$ is usually achieved by a weakly directional antenna (WDA). It must therefore have increased electrical strength for a given radiated power.
2. If the orthogonality condition is satisfied, the number of MBAA beams formed cannot exceed the number of radiators, and the step d between them is limited to overlap the area. These two factors determine the angular resolution of the radar.
3. Like any PAA, the MBAA has dispersive properties that limit the bandwidth of the probe signals (PS) used [4].

Some studies [7–9] have demonstrated the possibility to compensate or completely eliminate these disadvantages and limitations using the MIMO radar technology. The essence of this technology is as follows. A transmitting M-element AA radiates an M-component system of mutually orthogonal coherent PS. The width of partial DPs of this AA $\Delta\theta_{\text{el}}$ should be equal to $\Delta\theta_{\text{svy}}$. In turn, the PS orthogonality supports the assumption that the superposition of the received echo signals, after reflection from the target, can be divided into M independent channels with uncorrelated noise $\overline{n_i n_j} = \begin{cases} \sigma^2_{\text{noise0}} & \text{at } i = j, \\ 0 & \text{at } i \neq j, \end{cases}$ $i, j \in M$, where σ^2_{noise0} is the noise variance, assumed to be equal in all channels for simplicity, and the line above is the averaging symbol.

MIMO radars are limited to parallel space survey because there is virtually no PS interference on the target. In search mode, the radar is assumed to have a multi-beam pattern at the receiver, as in a traditional radar. We consider the implementation of parallel view in MIMO radars combined with spatiotemporal processing (SSP), which forms the optimal pre-threshold statistics (PTS) according to the Neyman–Pearson criterion [10].

This integrated approach allows MIMO radars to be compared with traditional radars at the PTS level in terms of providing equivalent detection quality index when searching for targets. Once this problem is solved, the design conditions and principles of MIMO radar can be determined to offer advantages over traditional radars despite possible practical difficulties.

The present work sets out to synthesize the SSP structure in a MIMO radar and compare it with similar processing in a traditional parallel-view radar assuming similar detection quality performance.

PARALLEL SURVEY AND TARGET DETECTION IN TRADITIONAL RADARS WITH MBAA

We consider the design principles of traditional radars with MBAA for parallel survey of a given angular sector $\Delta\theta_{\text{svy}}$. The most common is to use WDA for transmission and MBAA for reception, as shown in Fig. 1. Without loss of generality, it is assumed that the WDA DP width $\Delta\theta_{\text{WDA}} = \Delta\theta_{\text{svy}}$ and the number of independent beams of the linear MBAA is equal to the number of emitters N, arranged in equidistant steps $d_{\text{r}} \leq \dfrac{\lambda_0}{1+\left|\sin\left(\Delta\theta_{\text{svy}}/2\right)\right|}$, where λ_0 is the operating wavelength. In addition, the narrow-bandwidth condition of the excitation signal $u(t)$ with a bandwidth Δf_{s} [2, 8] is imposed on the receiving MBAA:

$$T_{\text{a}} = (N-1)d_{\text{r}}\sin\left(\frac{\Delta\theta_{\text{svy}}}{2}\right) \ll \frac{1}{\Delta f_{\text{s}}}, \qquad (1)$$

where T_{a} is the time of filling the MBAA aperture with a pulse of equivalent duration $\tau_{\text{p.e.}} = 1/\Delta f_{\text{s}}$.

We define the PS complex envelope $\dot{U}_{\text{ref}}(t)$ following reflection from a point target with a three-component vector of information parameters $\boldsymbol{\kappa} = \{R_{\text{TRG}}, V_{\text{rTRG}}, \theta_{\text{TRG}},\}$, where R_{TRG}, V_{rTRG}, and θ_{TRG} are range, radial velocity, and angular coordinate of the target, respectively. At the MBAA output, it has the following form:

$$\begin{aligned}&\dot{U}_{\text{ref}}(t,\boldsymbol{\kappa}) = \\ &= F_{\text{tr}}(\theta_{\text{TRG}})\dot{U}_{\text{tr}}(\gamma_{\text{TRG}}t - \tau_{\text{TRG}})\,\mathrm{e}^{j2\pi f_0\gamma_{\text{TRG}}t} = \\ &= U_{\text{tr}}(t'-\tau_{\text{TRG}})\,\mathrm{e}^{j2\pi f_0 t'},\end{aligned} \qquad (2)$$

where $F_{\text{tr}}(\theta_{\text{TRG}})$ is the WDA DP level in the θ_{TRG} direction; $\tau_{\text{TRG}} = 2R_{\text{TRG}}/\text{c}$ is signal delay time; c is the speed of light; $\gamma_{\text{TRG}} = 1 \pm 2V_{\text{r}}/\text{c}$ is a Doppler time scale change coefficient for which $F_{\text{D}} = 2V_{\text{r}}/\lambda_0$ is a Doppler frequency, $t' = \gamma_{\text{TRG}}t$.

It can be further assumed that $F_{\text{tr}}(\theta_{\text{TRG}}) \approx$ const and that the complex envelope of the received signal is decomposed by condition (1) into a scalar time function and an N-dimensional vector $\boldsymbol{\beta}(\theta_{\text{TRG}})$ of spatial phases in the single-target situation:

$$\begin{aligned}&U_{\text{r}}(t'-\theta_{\text{TRG}}) = U_{\text{tr}}(t'-\tau_{\text{TRG}})\boldsymbol{\beta}(\theta_{\text{TRG}}),\\ &\boldsymbol{\beta}(\theta_{\text{TRG}}) = \exp\left\{j\frac{2\pi d_{\text{r}}}{\lambda_0}(n-1)\sin\theta_{\text{TRG}}\right\}_{n=1}^{N}.\end{aligned} \qquad (3)$$

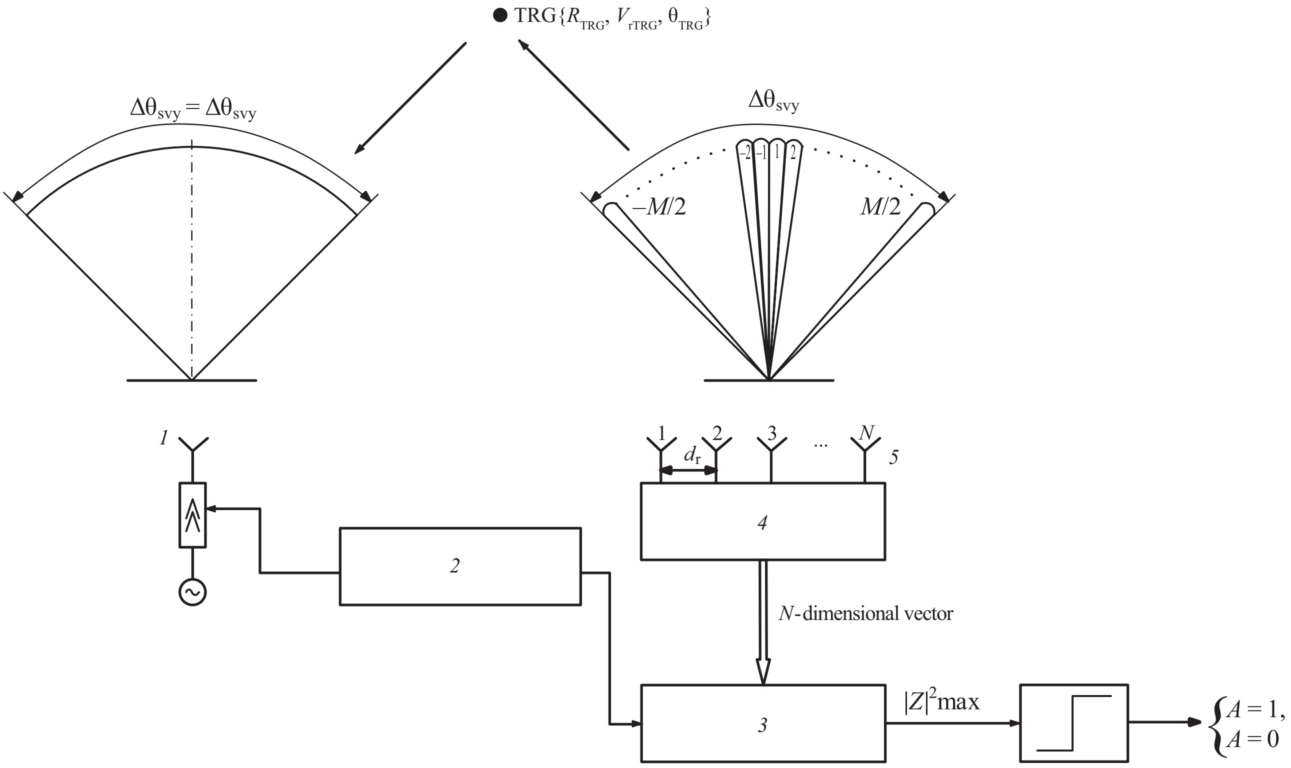

TRG{R_{TRG}, V_{rTRG}, θ_{TRG}}
$\Delta\theta_{\text{svy}} = \Delta\theta_{\text{svy}}$
$\Delta\theta_{\text{svy}}$
−2
−1
1
2
$-M/2$
$M/2$
1
1
2
3
...
N
5
d_{r}
4
2
N-dimensional vector
3
$|Z|^2$max
$A = 1,$
$A = 0$

Fig. 1. Traditional radar with parallel space survey: (*1*) WDA; (*2*) PS shaper; (*3*) temporal processing (PS shaping); (*4*) spatial processing (diagram-forming scheme); (*5*) MBAA. TRG{R_{TRG}, V_{rTRG}, θ_{TRG}} is target; *A* is the threshold at which a decision is made about the presence or absence of a signal

We consider the multichannel target detection problem in the classical mixture reception formulation, $y(t) = A\dot{U}_{\text{r}}(t) + n(t)$, where $A = \{1, 0\}$ depending on the presence/absence of the valid signal, and $n(t)$ is white Gaussian noise. In the absence of correlating external interference with the same white Gaussian noise intensity in the MBAA channel, $\sigma_i^2 = \sigma_j^2 = \sigma_0^2(\overline{i, j} \in M)$, while the $(N \times N)$ dimensional correlation matrix function of the interference is represented by the following equation:

$$\mathbf{\Phi}(t - s') = N_0 \mathbf{I} \delta(t - s), \tag{4}$$

where $\mathbf{I}$ is the unit diagonal matrix; N_0 is the noise power spectral density; δ is the Dirac delta function.

Under these conditions, the optimal SSP by Neyman–Pearson criterion can be reduced to calculating the PTS in the form of the squared modulus of the weight integral:

$$\xi = |Z|^2 = \frac{1}{N_0}\left|\int \dot{Y}_{\Sigma}(t')\dot{U}^*_{\text{rec}}(t')dt'\right|^2, \tag{5.1}$$

$$\dot{Y}_{\Sigma}(t') = \mathbf{Y}^{\text{T}}(t', \theta_{\text{TRG}})\boldsymbol{\beta}^*(\theta_{\text{TRG}i}), \tag{5.2}$$

where $\mathbf{Y}$ is the input vector.

In Eqs. (5.1) and (5.2), the symbols (*) and ($^\text{T}$) stand for complex conjugation and transposition, respectively. Equation (5.2) defines the complex amplitude at the output of the ith secondary MBAA channel phased in the expected direction $\theta_{\text{TRG}i}$ ($i \in \overline{1, N}$). As mentioned above, the N rays formed by MBAA should cover the entire given survey sector $\Delta\theta_{\text{svy}}$. In practice, the lossless formation of orthogonal rays can be conveniently implemented based on the fast Fourier transform (FFT) algorithm when $N = 2^q$, where q is an integer (the Butler matrix in analogue form [1]). The FFT algorithm converts the counts of the N-dimensional vector of the input signal (3) (primary MBAA channels) into a vector of N orthogonal rays (secondary channels (DP)) using the $(N \times N)$ transformation matrix $\mathbf{W}(\theta) = \{w\}_{n,i}^{N,N}$, where $w_{ni} = \exp\left(j2\pi n/N\right)$; n, i are the numbers of the primary and secondary MBAA channels, respectively.

After implementing the FFT, Eq. (5.1) can be considered as expressing the PTS in each of the N secondary channels if we assume: $\dot{Y}_{\Sigma}(t') = \mathbf{Y}^{\text{T}}(t', \theta_{\text{TRG}})\mathbf{W}^*(\theta_{\text{TRG}i})$, where $\theta_{\text{TRG}i} = \arcsin\left[(i - 1/2)\lambda_0/(N-1)d_{\text{r}}\right]$ (see Fig. 1).

The principles of the parallel survey described above are common to the traditional radar systems with MBAA. They have some features that are more important for comparison with MIMO radars. These include:

1. The target is simultaneously irradiated by a single coherent PS $\dot{U}_{\text{tr}}(t)$ at the carrier frequency f_0 with a given average power.
2. Increasing the number of elements in the WDA, e.g., in the form of a small AA, is often impossible in principle. This is because it is accompanied by a narrowing of $\Delta\theta_{\text{WDA}}$, which does not effectively illuminate the search area $\Delta\theta_{\text{svy}}$.
3. The resolving power of MBAA beams is determined by the geometric size of their aperture $L_{\text{multiAA}} = (N - 1)d_{\text{r}}$. In this case, according to point 2, the number of orthogonal beams is limited by the number of primary channels N, and the increase in step d_{r} is limited by the width of the specified survey sector $\Delta\theta_{\text{svy}}$. These factors do not allow the MBAA beams to be narrowed or their number to be increased.
4. The design of a traditional parallel space survey radar (Fig. 1) implements a factorized representation of the PTS (3). This allows the sequential SSP to be divided into spatial (DP) and spatiotemporal processing (Woodworth function) in this order.

It should be noted here that the reverse order is impractical as it would require the same temporal accumulation to be performed in each primary MBAA channel prior to the spatial accumulation, which conveniently performed only once.

The above features and limitations are rare. They are mostly removed in MIMO radars due to the increased dimensionality of the problem. Instead of a single PS, N orthogonal but coherent signals are transmitted simultaneously, providing additional freedom for radar surveillance.

Assuming equivalent PTS and detection performance, we now turn to the analysis of the differences between MIMO radars and traditional radars.

PARALLEL SURVEY AND TARGET DETECTION IN MIMO RADARS

We start with the main characteristic of MIMO radars, which is the illumination of the coverage area by a system of orthogonal arrays that excite the AA with step d_{r}, shown in the left part of Fig. 2. The array generally emits a vector signal with a complex envelope $\dot{\mathbf{U}}_{\text{tr}}(t) = \{\dot{U}_{\text{tr}\,m}\}_{m=1}^{M}$. The orthogonality condition of these components should be satisfied for all directions θ within the sector $\Delta\theta_{\text{svy}}$. Then the normalized correlation coefficient between the pth and qth components can be described by the following equation:

$$\rho_{pq} = \\ = \frac{\int \dot{U}_{p\text{tr}}\left[t + (p-1)\frac{d_{\text{tr}}}{\text{c}}\sin\theta\right]dt \int \dot{U}^*_{q\text{tr}}\left[t + (q-1)\frac{d_{\text{tr}}}{\text{c}}\sin\theta\right]dt}{\left[\int |\dot{U}_{p\text{tr}}(t)|^2\, dt \int |\dot{U}_{q\text{tr}}(t)|^2\, dt\right]^{1/2}}. \tag{6}$$

The complex envelope of the total signal reaching the target has the following form:

$$\dot{\mathbf{U}}_{\text{tr}}(t,\theta_{\text{TRG}}) = \\ = \sum_{m=1}^{M} \dot{U}_{m\text{tr}}\left[t+(m-1)\frac{d_{\text{tr}}}{\text{c}}\right]\exp\left[j\frac{2\pi d_{\text{tr}}}{\lambda_0}(m-1)\sin\theta_{\text{TRG}}\right]. \quad (7)$$

As in the previous case, it is assumed to be narrowband. This satisfies condition (1). Note that for MIMO radars, by operating on orthogonal signals, e.g., separated by Δf_0 carrier frequencies [11, 12], the traditional limitations of the AA bandwidth are practically eliminated. When received, they can be band-separated and thus their interference can be neglected. For the correctness of the comparison, it is assumed that the signal bandwidth transmitted by the MIMO radar AA is the same as that of a traditional radar and is equal to $\Delta f_{\text{s}} = (N-1)\Delta f_0$ under the condition $\Delta f_{\text{s}} \ll \Delta f_0$.

The average power of the vector signal (7) reaching the target can be represented by quadratic Hermite:

$$P_{\text{tr}} = \sum_{p=1}^{M}\sum_{q=1}^{M}\overline{\dot{U}_p\dot{U}_q}\exp\left[j\frac{2\pi d_{\text{tr}}}{\lambda_0}(p-q)\sin\theta_{\text{TRG}}\right] = \\ = \alpha^{\text{T}}(\theta_{\text{TRG}})\mathbf{I}\alpha^{*}(\theta_{\text{TRG}}). \quad (8)$$

We assume that the power (8) is the same as in the previous case, i.e., $P_{\text{tr}} = P_{\text{t.p.}}G_{\text{WDA}}$, where $P_{\text{t.p.}}$ is the transmitter power at the WDA input of the traditional radar and G_{WDA} is the directional coefficient of the WDA.

For a point target with a vector of information parameters $\boldsymbol{\kappa} = \{t', V_{\text{r TRG}}, \theta_{\text{TRG}}\}$, the complex amplitude of the reflected signal is as follows:

$$\dot{\mathbf{U}}_{\text{ref}}(t',\tau_{\text{TRG}},\theta_{\text{TRG}}) = F_{\text{tr}}(\theta_{\text{TRG}})\sum_{m=1}^{M}\dot{U}_{m\text{tr}}(t'-\tau_{\text{TRG}})\times \\ \times\exp\left[j\frac{2\pi d_{\text{tr}}}{\lambda_0}(m-1)\sin\theta_{\text{TRG}}\right]. \quad (9)$$

We assume that all radiators of the transmitting AA have the same weakly directional DP adapted to the coverage area, i.e., $F_{\text{tr}}(\theta_{\text{TRG}}) = \text{const}$ for all $\theta_{\text{TRG}} \in \Delta\theta_{\text{svy}}$. Then, Eq. (9) can be viewed as the M-dimensional vector $\dot{\mathbf{U}}_{\text{ref}}(t) = \left\{\dot{U}_{m\text{tr}}(t,\tau'_{\text{TRG}})\text{e}^{j\alpha_m(\theta_{\text{TRG}})}\right\}_{m=1}^{M}$, which coincides with the vector exciting the single receiving WDA in the right part of Fig. 2, where $\alpha_m(\theta_{\text{TRG}})$ is the phase run-up of the mth partial signal.

We proceed with the synthesis of the SSP at the output of this WDA, given that the received signal $\dot{U}_{\text{rec}}(t) = \dot{U}_{\text{ref}}(t)$ is factorized into M temporal and spatial multipliers. We therefore divide it into M independent channels, as shown in Fig. 3:

$$\dot{\mathbf{U}}_{\text{rec}}(t',\tau_{\text{TRG}},\theta_{\text{TRG}}) = \mathbf{E}^{\text{T}}\otimes\dot{\mathbf{U}}_{\text{ref}}(t',\tau_{\text{TRG}},\theta_{\text{TRG}}) = \\ = \left\{\dot{U}_{m\text{tr}}(t'-\tau_{\text{TRG}})\text{e}^{j\alpha_m(\theta_{\text{TRG}})}\right\}_{m=1}^{M}, \quad (10)$$

where $\mathbf{E} = \{1\}_{m=1}^{M}$ is the M-dimensional unit vector and $\otimes$ is the Kronecker product symbol.

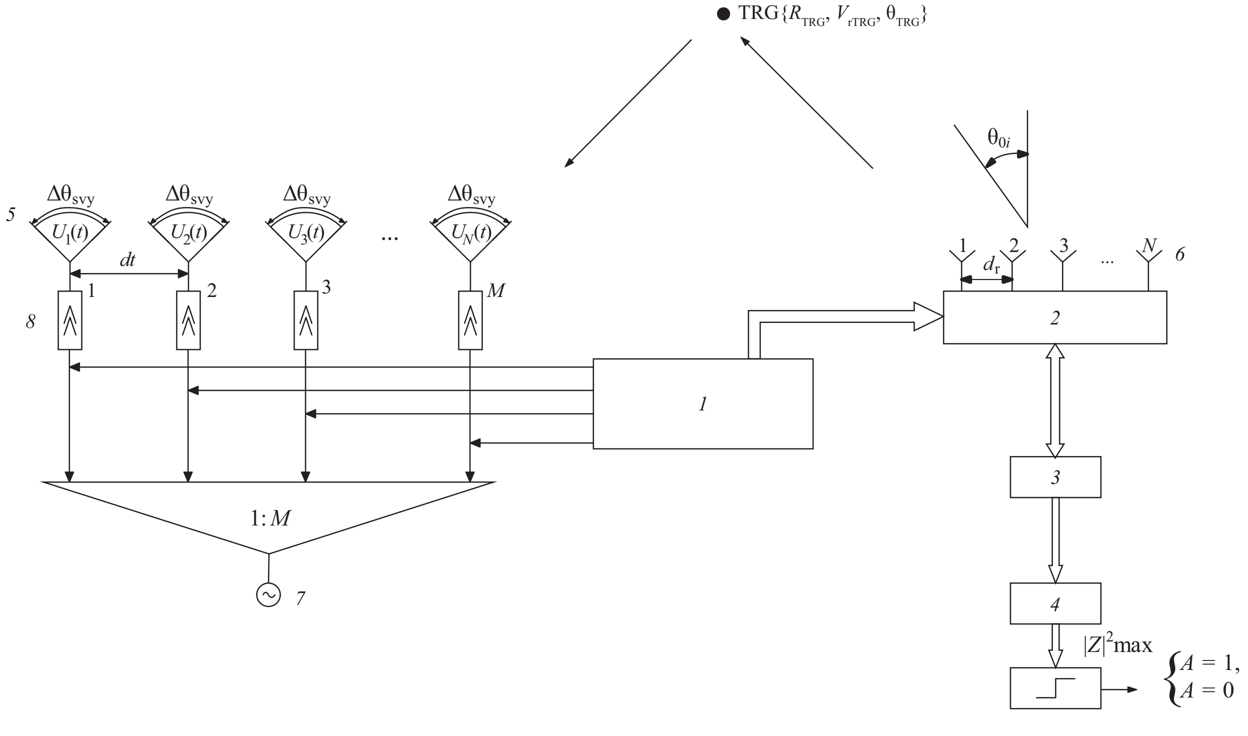

TRG{R_{TRG}, V_{rTRG}, θ_{TRG}}
θ_{0i}
$\Delta\theta_{\text{svy}}$
$\Delta\theta_{\text{svy}}$
$\Delta\theta_{\text{svy}}$
$\Delta\theta_{\text{svy}}$
5
$U_1(t)$
$U_2(t)$
$U_3(t)$
...
$U_N(t)$
dt
1
2
3
M
8
1
2
3
4
1:M
7
1
2
3
...
N
6
d_{r}
$|Z|^2$max
A = 1,
A = 0

Fig. 2. MIMO radar with parallel space survey: (*1*) orthogonal PS generator; (*2*) temporal processing; (*3*) spatial processing; (*4*) PTS generator; (*5*) transmitting AA; (*6*) receiving AA; (*7*) driving generator; (*8*) power amplifier

With respect to the noise properties, these channels are independent and their correlation matrix function is described by Eq. (4), which applies here as well. In the white Gaussian noise background, optimal temporal processing is implemented by a set of matched filters (MFs) for each partial signal $U_m(t)$ in the receive channel. However, in contrast to (5), in the one-target situation for the expected direction $\theta_{\mathrm{TRG}i}$ in the expression for PTS, the time multiplier of the signal is a vector rather than a scalar and is of the following form:

$$\xi = |Z|^2 = \mathbf{Y}^{\mathrm{T}}(t',\theta_{\mathrm{TRG}})\mathbf{U}^*(t',\theta_i) = \\ = \frac{1}{N_0}\left|\sum_{m=1}^{M}\dot{Y}_m(t',\theta_{\mathrm{TRG}})\dot{U}^*_{0m}(t')\mathrm{e}^{-j\frac{2\pi d_{\mathrm{tr}}}{\lambda_0}(\sin\theta_{\mathrm{TRG}}-\sin\theta_i)}\right|^2 = \\ = \frac{1}{N_0}\left|\sum_{m=1}^{M}Y_{m\,\mathrm{mf}}(0)\mathrm{e}^{-j\frac{2\pi n d_{\mathrm{tr}}}{\lambda_0}(m-1)(\sin\theta_{\mathrm{TRG}}-\sin\theta_i)}\right|^2, \qquad (11)$$

where $\dot{U}_{0m}(t')$ is the complex envelope of the expected signal in the mth partial channel, $Y_{m\,\mathrm{mf}}(0)$ is the result of temporal accumulation of the signal, which corresponds to the maximum amplitude at $t' = 0$ at the MF output in the mth channel.

An important feature of the scheme in Fig. 3 is that temporal accumulation precedes spatial accumulation, indicating an inverse order relative to traditional SSP.

Using $t' \neq 0$ in (11), the output signal $\dot{Y}_{m\,\mathrm{mf}}(t')$ can be considered as a frequency-temporal mismatch in range and velocity, and the formula generally as a multidimensional mismatch function in the range-velocity-angle coordinates, similar to [13].

An additional interpretation of Eq. (11) can be provided. It corresponds to the traditional AA multiplier formula, where the vector $\dot{\mathbf{Y}}_{\mathrm{mf}}(t')$ has the form of a time-dependent amplitude distribution of M-element equidistant AA with step d_{tr}. If $Y_{\mathrm{mf}}(0)$ = const in all channels, then this equivalent distribution is uniform at the time corresponding to the signal maximum. In antenna theory, this transformation of a single receive radiator into an AA identical to the transmit one (Fig. 4) corresponds to the concept of the synthesized aperture. However, most papers dealing with MIMO radars use the term virtual sublattice/lattice [7, 14, 15].

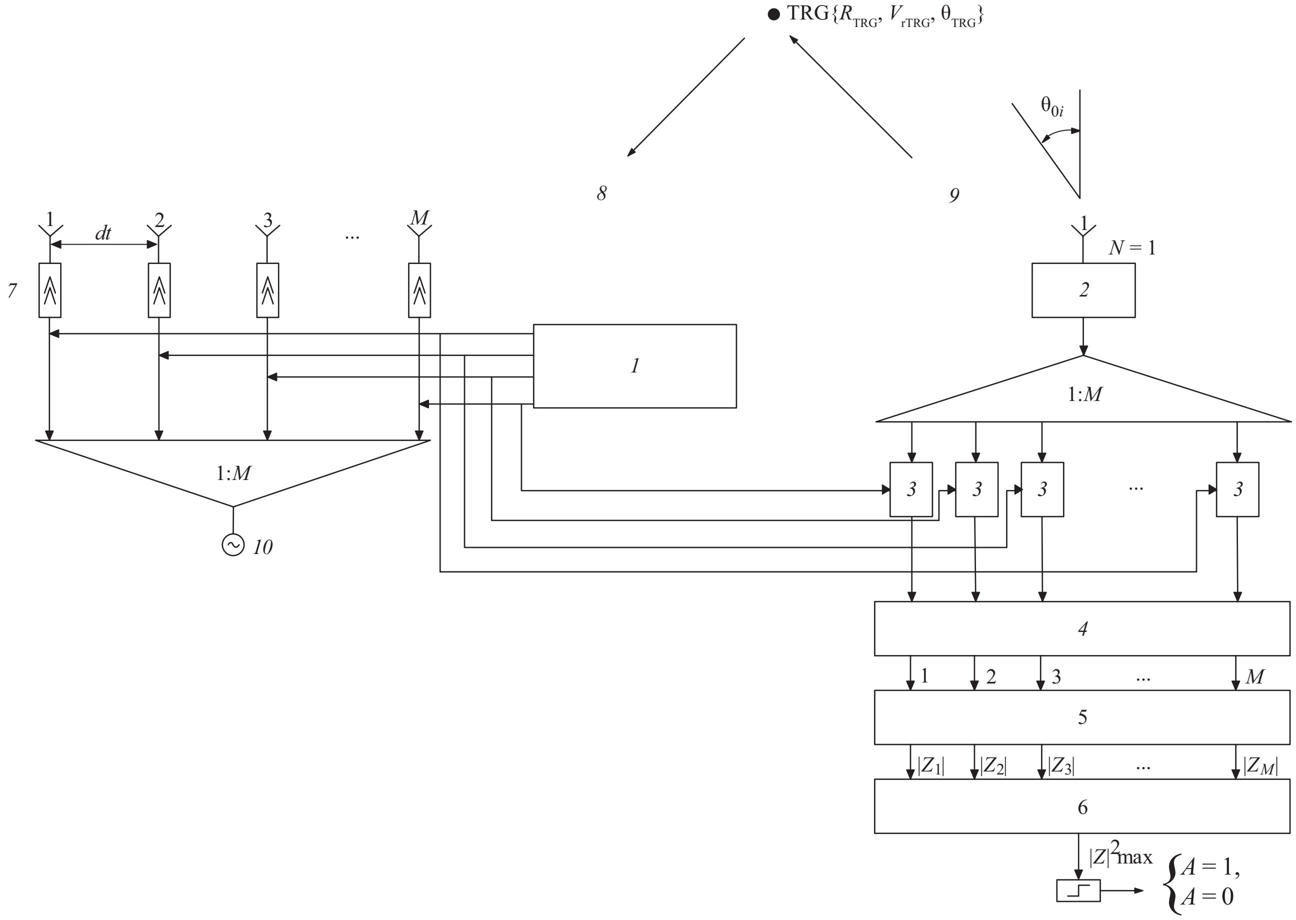


Fig. 3. Sequence of SSP steps in MIMO radars: (*1*) orthogonal PS shaper; (*2*) low-noise amplifier; (*3*) MF; (*4*) diagram-forming scheme (*M*-point FFT); (*5*) PTS shaper; (*6*) maximum sampling; (*7*) power amplifier; (*8*) transmitting AA; (*9*) receiving AA; (*10*) driving generator

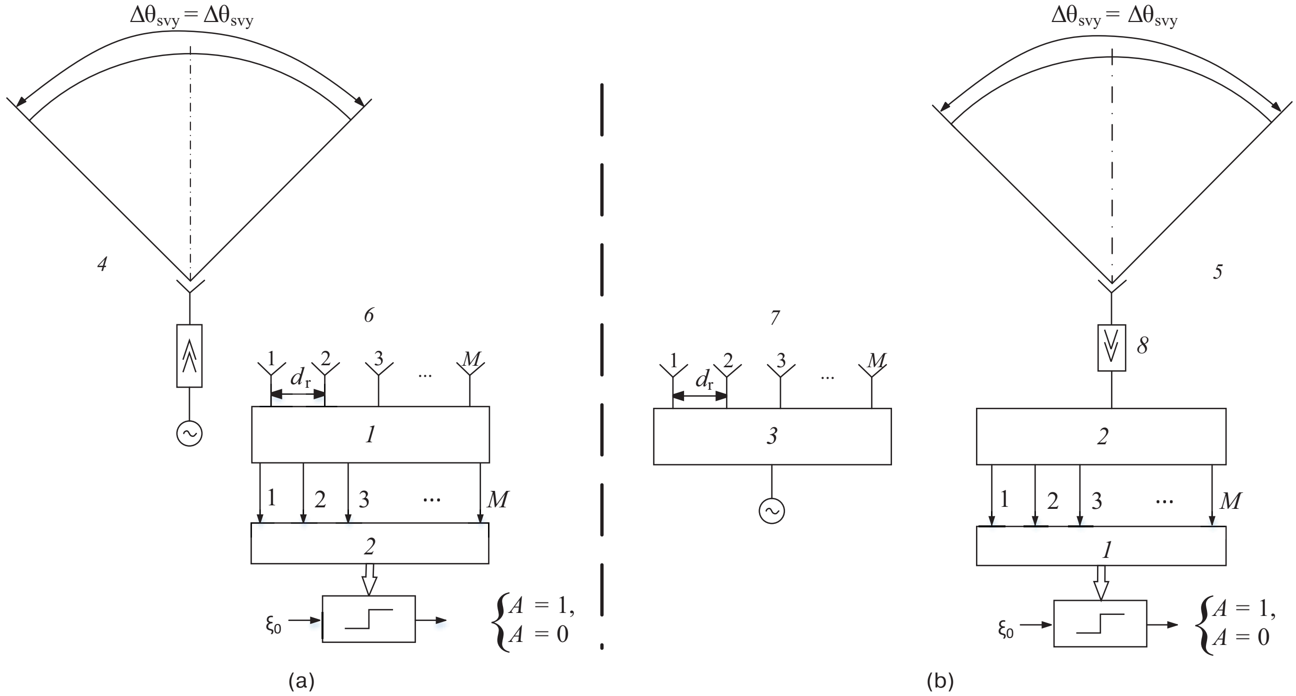


Fig. 4. Mirror construction of traditional (a) and MIMO (b) radars with parallel space survey: (*1*) diagram-forming scheme; (*2*) temporal processing; (*3*) orthogonal PS shaper; (*4*) transmitting WDA; (*5*) receiving WDA; (*6*) receiving MBAA; (*7*) transmitting AA; (*8*) low-noise amplifier

Due to the a priori uncertainty of the angular position of the point target ($\theta_{\mathrm{TRG}i}$) after time accumulation, it is necessary to implement a multi-beam diagram-forming scheme at the virtual sublattice output. As shown above, a digital FFT algorithm can be used to form orthogonal beams and thus to obtain M secondary receive beams:

$$|Z|^2 = \max|\mathbf{Y}_{\mathrm{mf}}^{\mathrm{T}}(0)\mathbf{W}^*(\theta_{0i})|^2,$$
$$\mathbf{W}(\theta_{0i}) = \left\{\exp\left(j\frac{2\pi mi}{M}\right)\right\}_{m,i=-\frac{M}{2}}^{\frac{M}{2}}. \tag{12}$$

From the obtained structure of the MIMO radar, which performs parallel space survey, it is possible to infer its mirror structure compared to the traditional radar, as shown in Fig. 4.

In Fig. 4, the channels for the Doppler processing at the inputs of the threshold devices are not shown in the versions compared. In both cases, the Doppler filtering systems are identical and can be implemented by different methods in the form of inter-cycle compensation or a set of filters tuned to the expected radial velocities $V_{\mathrm{r}0} = \dfrac{\lambda\Delta\varphi_0}{2\Delta T}$, where $\Delta\varphi_0$ is the phase difference between adjacent packet pulses and ΔT is the repetition interval. This means that the informative parameter *velocity* has no specific characteristics for the narrowband MIMO radar. Under the chosen conditions, PTS equivalence of both versions ensures identity of their statistical detection quality index.

The main characteristics of MIMO radars with transmitting AA and receiving single channel arrangement (WDA) are as follows:

1. The target is simultaneously irradiated by a vector of coherent orthogonal PS $\dot{\mathbf{U}}_{\mathrm{tr}}(t)$ emitted, for example, by AR elements at different carrier frequencies f_{0m}. Increasing the number of AA radiators not only does not lead to a narrowing of the survey sector $\Delta\theta_{\mathrm{svy}}$ but also reduces the requirements on the electrical strength of the path for a fixed radiated power.
2. The angular resolution in this MIMO radar is only determined by the size of the transmitting AA $L_{\mathrm{tr}} = (M - 1)d_{\mathrm{tr}}$ into which the received WDA is transformed by the initial frequency-temporal processing into a virtual sublattice.
3. As in the traditional case of a single WDA for reception, the SSP is factorized into temporal and spatial variants. However, their order is not fundamental and for practical purposes can be reversed compared to the MBAA version.
4. The advantages of MIMO radars mentioned in point 1 are partially offset by certain difficulties in the practical implementation of the transmitting AA. These difficulties include the need to form a set of coherent signals and maintain their coherence when routing ultrahigh frequency signals through the channels of the transmitting AA.

The coherent spatiotemporal processing in the virtual sublattice channels is performed at a lower level of the valid signal compared to the traditional version, where MFs of the secondary channels operate following their coherent spatial accumulation. However, the detection quality is theoretically the same for the same SNR value while maintaining linearity.

CONCLUSIONS

The results obtained in the paper allow the following conclusions:

1. The comparison between the radar of traditional construction and the MIMO radar with a single channel per reception in the parallel target search mode shows the equivalence of their statistical PCOs under the following conditions:
- same resulting SNR and PTS in the detector;
- same bandwidth of coherent PSs and linearity in receive;
- same survey sectors $\Delta\theta_{svy}$;
- the possibility to divide the SSP into frequency-temporal and spatial components.
2. The theoretical equivalence of the versions compared is achieved by mirroring their structural schemes (see Fig. 4).

Traditional version:
- WDA for transmission and *M*-element MBAA for reception;
- spatial accumulation precedes temporal accumulation.

MIMO radar:
- *M*-element AA for transmission and WDA for reception;
- temporal accumulation precedes spatial accumulation.

The total number of SSP channels is equal. When the single WDA for reception is replaced by a multichannel receiving AA, the main advantages of MIMO radars over traditional radars become apparent. This more complex case will be discussed in the next paper.

Authors' contributions

B.M. Vovshin—general statement of the problem, development of principles for constructing MIMO radars, development of a methodology for comparing MIMO radars and traditional parallel-survey radars.

A.A. Pushkov—development of structural design schemes, derivation of analytical ratios, analysis of the results of comparison of radar design options.

E.M. Khalturina—development of structural construction schemes, derivation of analytical ratios, writing text, design of material.

REFERENCES

1. Chernyak V.S. About new and old ideas in radar: MIMO radars. *Uspekhi sovremennoi radioelektronik = Achievements of Modern Radioelectronics*. 2011;2:5–20 (in Russ.).
2. Li J., Stoica P. *MIMO Radar Signal Processing*. New York: Wiley; 2009. 448 p.
3. Bakhrakh L.D., Voskresenskii D.I. (Eds.) *Problemy antennoi tekhniki* (*Antenna Technology Problems*). Moscow: Radio i svyaz'; 1989. 368 p. (in Russ.).
4. Mailloux R.J. *Phased Array Antenna Handbook*. 2nd Ed. London: Artech House; 2005. 479 p.
5. Chernyak V.S. Signal detection with MIMO radars. *Uspekhi sovremennoi radioelektronik = Achievements of Modern Radioelectronics*. 2014;7:35–48 (in Russ.).
6. Vovshin B.M. The intertialless air surveillance by ultra wideband radar signals. *Antenny*. 2006;7(102):92–100 (in Russ.).
7. Bergin J., Guerci J.R. *MIMO Radar: Theory and Application*. London: Artech House; 2018. 221 p.
8. Dorey J., Garnier G., Auvray G. RIAS Radar a impulsion et Antenne Synthetique. In: *Proceedings Colloque International sur le Radar*. Paris. April. 1989. P. 112–115 (in French).
9. Vovshin B.M. Parallel Surveillance Ultra-wide-band Radars with Orthogonal Ranging Signals. In: *Proceedings on International Radar Symposium* (*IRS-2007*). Cologue, Germany. P. 461–466.
10. Shirman Ya.D. (Ed.). *Radioelektronnye sistemy. Osnovy postroeniya i teoriya* (*Radioelectronic systems. Fundamentals of Construction and Theory*). Moscow: Radiotekhnika; 2007. 512 p. (in Russ.).
11. Vovshin B.M., Immoreev I.Y. Influence of dispersion properties phased array antenna on the signal to noise ratio in radars with broadband signals. *Radiotekhnika*. 1985;7:74–92 (in Russ.).
12. Kalinin V.I., Chapursky V.V., Cherepenin V.A. Super-Resolution of Radar and Radio Holography Systems Based on a MIMO Retrodirective Antenna Array. *J. Commun. Technol. Electron.* 2021;66(6):727–736. https://doi.org/10.1134/S1064226921060139
    [Original Russian Text: Kalinin V.I., Chapursky V.V., Cherepenin V.A. Super-Resolution of Radar and Radio Holography Systems Based on a MIMO Retrodirective Antenna Array. *Radiotekhnika i elektronika*. 2021;66(6):614–624 (in Russ.). https://doi.org/10.31857/S0033849421060139]
13. Chapurskii V.V., Slukin G.P., Filatov A.A., Koroteev D.E. Virtual MIMO radars and their comparison based on generalized uncertainty functions. *Uspekhi sovremennoi radioelektroniki = Achievements of Modern Radioelectronics*. 2023;77(5):5–19 (in Russ.). https://doi.org/10.18127/j20700784-202305-01

14. Brookner E. MIMO Radars and their Conventional Equivalents. In: *Proc. on IEEE International Radar conference*. May 10–15, 2015, Arlington, VA, USA, P. 918–924. https://doi.org/10.1109/RADAR.2015.7131126
15. Tebaldini S., Manzoni M., Ferro-Famil L., Banda F., Giudici D. FDM MIMO SAR Tomography. In: EUSAR 2024 – *15th European Conference on Synthetic Aperture Radar*. Munich Germany. 2024. P. 709–714.

СПИСОК ЛИТЕРАТУРЫ

1. Черняк В.С. О новых и старых идеях в радиолокации: MIMO РЛС. *Успехи современной радиоэлектроники*. 2011;2:5–20.
2. Li J., Stoica P. *MIMO Radar Signal Processing*. New York: Wiley; 2009. 448 p.
3. *Проблемы антенной техники*; под ред. Л.Д. Бахраха и Д.И. Воскресенского. М.: Радио и связь; 1989. 368 с.
4. Mailloux R.J. *Phased Array Antenna Handbook.* 2nd Ed. London: Artech House; 2005. 479 p.
5. Черняк В.С. Обнаружение сигналов в MIMO РЛС. *Успехи современной радиоэлектроники*. 2014;7:35–48.
6. Вовшин Б.М. Безынерционный обзор пространства сверхширокополосными радиолокационными сигналами. *Антенны*. 2006;7(102):92–100.
7. Bergin J., Guerci J.R. *MIMO Radar: Theory and Application*. London: Artech House; 2018. 221 p.
8. Dorey J., Garnier G., Auvray G. RIAS Radar a impulsion et Antenne Synthetique. In: *Proceedings Colloque International sur le Radar*. Paris. April. 1989. P. 112–115 (in French).
9. Vovshin B.M. Parallel Surveillance Ultra-wide-band Radars with Orthogonal Ranging Signals. In: *Proceedings on International Radar Symposium* (*IRS-2007*). Cologue, Germany. P. 461–466.
10. *Радиоэлектронные системы. Основы построения и теория;* под ред. Я.Д. Ширмана. М.: Радиотехника; 2007. 512 с.
11. Вовшин Б.М., Иммореев И.Я. Влияние дисперсионных свойств фазированной антенной решетки на отношение сигнал/шум в РЛС с широкополосными сигналами. *Радиотехника*. 1985;7:74–92.
12. Калинин В.И., Чапурский В.В., Черепенин В.А. Сверхразрешение в системах радиолокации и радиоголографии на основе MIMO антенных решеток с рециркуляцией сигналов. *Радиотехника и электроника*. 2021;66(6):614–624. https://doi.org/10.31857/S0033849421060139
13. Чапурский В.В., Слукин Г.П., Филатов А.А., Коротеев Д.Е. Виртуальное MIMO РЛС и их сравнение на основе обобщенных функций неопределенности. *Успехи современной радиоэлектроники*. 2023;77(5):5–19. https://doi.org/10.18127/j20700784-202305-01
14. Brookner E. MIMO Radars and their Conventional Equivalents. In: *Proc. on IEEE International Radar conference*. May 10–15, 2015, Arlington, VA, USA: P. 918–924. https://doi.org/10.1109/RADAR.2015.7131126
15. Tebaldini S., Manzoni M., Ferro-Famil L., Banda F., Giudici D. FDM MIMO SAR Tomography. In: EUSAR 2024 – *15th European Conference on Synthetic Aperture Radar*. Munich, Germany: 2024. P. 709–714.

About the Authors

Boris M. Vovshin, Dr. Sci. (Eng.), Professor, Department of Radio Wave Processes and Technology, Institute of Radio Electronics and Informatics, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia); Chief Researcher, PJSC “ALMAZ R&P Corp.” (110, Dmitrovskoe sh., Moscow, 127411 Russia). E-mail: boris@eleron.net. Scopus Author ID 7801355022, https://orcid.org/0009-0003-1357-5866

Alexander A. Pushkov, Cand. Sci. (Eng.), Head of the Scientific and Technical Center, PJSC “ALMAZ R&P Corp.” (110, Dmitrovskoe sh., Moscow, 127411 Russia). E-mail: aapushkov@mail.ru. https://orcid.org/0009-0005-3831-4528

Elizaveta M. Khalturina, Engineer, PJSC “ALMAZ R&P Corp.” (110, Dmitrovskoe sh., Moscow, 127411 Russia). E-mail: e.m.halturina@yandex.ru. https://orcid.org/0009-0003-8527-419X

Об авторах

Вовшин Борис Михайлович, д.т.н., профессор, кафедра радиоволновых процессов и технологий, Институт радиоэлектроники и информатики, ФГБОУ ВО «МИРЭА – Российский технологический университет (119454, Россия, Москва, пр-т Вернадского, д. 78); главный научный сотрудник, ПАО «Научно-производственное объединение «Алмаз» имени академика А.А. Расплетина (ПАО «НПО «Алмаз») (127411, Россия, Москва, Дмитровское шоссе, д. 110). E-mail: boris@eleron.net. Scopus Author ID 7801355022, https://orcid.org/0009-0003-1357-5866

Пушков Александр Александрович, к.т.н., начальник научно-технического центра, ПАО «Научно-производственное объединение «Алмаз» имени академика А.А. Расплетина (ПАО «НПО «Алмаз») (127411, Россия, Москва, Дмитровское шоссе, д. 110). E-mail: aapushkov@mail.ru. Scopus Author ID 54407412700, https://orcid.org/0009-0005-3831-4528

Халтурина Елизавета Максимовна, инженер 3 категории, ПАО «Научно-производственное объединение «Алмаз» имени академика А.А. Расплетина (ПАО «НПО «Алмаз») (127411, Россия, Москва, Дмитровское шоссе, д. 110). E-mail: e.m.halturina@yandex.ru. https://orcid.org/0009-0003-8527-419X

Translated from Russian into English by K. Nazarov
Edited for English language and spelling by Thomas A. Beavitt

Modern radio engineering and telecommunication systems

Современные радиотехнические и телекоммуникационные системы

UDC 621.372.85
https://doi.org/10.32362/2500-316X-2025-13-3-84-91
EDN NHBOSK

cc BY

SHORT COMMUNICATION

Control of the frequency response of a narrow-band filter for the X-band frequency based on a photonic crystal with a movable cylindrical defect

Evgeny A. Ryabov [@],
Anton A. Andreev,
Sergey A. Sergeev,
Alexander I. Mikhailov

Saratov National Research State University, Saratov, 410012 Russia
[@] Corresponding author, e-mail: k1u2r3ka@mail.ru

• **Submitted:** 12.11.2024 • **Revised:** 14.02.2025 • **Accepted:** 21.03.2025

Abstract

Objectives. The work set out to investigate the possibility and effectivity of using a movable cylindrical defect with metal pins in the design of a photonic crystal to control the frequency response of a narrow-band filter in a rectangular waveguide having a cross-section of 23 × 10 mm in the X-band, as well as to determine the most effective methods for controlling frequency response.

Methods. A numerical simulation of the frequency response of the filter was carried out using the *openEMS* software package, which is based on Maxwell's equations solved by the finite-difference time-domain method. The frequency response of the currently proposed and implemented filter construction in the X-band was further investigated in an experimental study.

Results. Numerical simulation shows that a resonant transmission peak in the stopband of the frequency response can be caused to appear by introducing a movable cylindrical defect having two metal pins into the center of a photonic crystal structure. In addition, the position of this peak on the frequency response can be effectively controlled by rotating the cylindrical defect around its axis. If the position of the defect remains unchanged, an increase in the frequency of the transmission peak occurs as a result of decreasing the period of the photonic crystal. However, the frequency of this resonant transmission peak is most strongly influenced by changes in the size of holes in the photonic structure. These changes can be used to control both the position and shape of the transmission peak, as well as the overall frequency response. At the same time, the difference in transmission remains practically unchanged when the cylinder rotates around its axis. The simulation results were confirmed by the data of an experimental study of the frequency response of photonic crystals made from PETG plastic using 3D printing technology.

Conclusions. The proposed, designed, and manufactured experimental samples of narrow-band filters in the X-band based on a photonic crystal demonstrated reliably variable transmission values and the possibility of controlling the resonant peak frequency and thus the entire frequency response, including operational control. This makes them very promising for practical use in radio-electronic equipment.

Keywords: narrow-band filter, resonant filter, microwave range, photonic crystal, 3D printing, openEMS

© E.A. Ryabov, A.A. Andreev, S.A. Sergeev, A.I. Mikhailov, 2025

For citation: Ryabov E.A., Andreev A.A., Sergeev S.A., Mikhailov A.I. Control of the frequency response of a narrow-band filter for the X-band frequency based on a photonic crystal with a movable cylindrical defect. *Russian Technological Journal.* 2025;13(3):84–91. https://doi.org/10.32362/2500-316X-2025-13-3-84-91, https://www.elibrary.ru/NHBOSK

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

КРАТКОЕ СООБЩЕНИЕ

Управление амплитудно-частотной характеристикой узкополосного фильтра для X-диапазона частот на основе фотонного кристалла с подвижным цилиндрическим дефектом

Е.А. Рябов @**,**
А.А. Андреев,
С.А. Сергеев,
А.И. Михайлов

Саратовский национальный исследовательский государственный университет имени Н.Г. Чернышевского, Саратов, 410012 Россия
@ *Автор для переписки, e-mail: k1u2r3ka@mail.ru*

• **Поступила:** 12.11.2024 • **Доработана:** 14.02.2025 • **Принята к опубликованию:** 21.03.2025

Резюме

Цели. Цель работы – исследовать возможность и эффективность использования в конструкции фотонного кристалла подвижного цилиндрического дефекта с металлическими штырями для управления амплитудно-частотной характеристикой (АЧХ) узкополосного фильтра на прямоугольном волноводе с сечением 23 × 10 мм в трехсантиметровом диапазоне (X-диапазоне), определить способы наиболее эффективного управления АЧХ.

Методы. Для численного моделирования АЧХ фильтра используется программный пакет *openEMS*, в основе которого лежит система уравнений Максвелла, решаемая методом конечных разностей во временной области. Проведено также экспериментальное исследование АЧХ действующего макета предложенной и созданной конструкции фильтра в трехсантиметровом диапазоне (X-диапазоне).

Результаты. Результаты численного моделирования показывают, что введение в центр конструкции фотонного кристалла подвижного цилиндрического дефекта с двумя металлическими штырями приводит к появлению в полосе запирания на АЧХ фильтра резонансного пика пропускания, положение которого эффективно управляется поворотом цилиндрического дефекта вокруг его оси. При неизменном положении цилиндрического дефекта уменьшение периода фотонного кристалла приводит к увеличению частоты пика пропускания. На частоту резонансного пика пропускания наиболее сильное влияние оказывает изменение размера отверстий в конструкции фотонного кристалла, что может использоваться как эффективный фактор для управления положением пика пропускания и формой всей АЧХ; при этом значение коэффициента пропускания при повороте цилиндрического дефекта вокруг его оси практически не изменяется. Проведены также экспериментальные исследования АЧХ фотонных кристаллов, изготовленных с использованием технологии 3D-печати из пластика PETG (полиэтилентерефталатгликоль), данные которых согласуются с результатами моделирования.

Выводы. Предложенные спроектированные и изготовленные экспериментальные модели узкополосных фильтров в трехсантиметровом диапазоне (X-диапазоне) на основе фотонного кристалла показали достаточные для практики изменения значения коэффициента пропускания и возможности эффективного управления частотой резонансного пика и всей формой АЧХ, что делает их весьма перспективными для практических применений в радиоэлектронной аппаратуре.

Ключевые слова: узкополосный фильтр, резонансный фильтр, СВЧ-диапазон, фотонный кристалл, 3D-печать, openEMS

Для цитирования: Рябов Е.А., Андреев А.А., Сергеев С.А., Михайлов А.И. Управление амплитудно-частотной характеристикой узкополосного фильтра для X-диапазона частот на основе фотонного кристалла с подвижным цилиндрическим дефектом. *Russian Technological Journal*. 2025;13(3):84–91. https://doi.org/10.32362/2500-316X-2025-13-3-84-91, https://www.elibrary.ru/NHBOSK

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

Electromagnetic waves are widely used in diverse fields of science and technology, including radiolocation and navigation services, as well as information and telecommunications technologies, medical equipment, etc. One of the most common types of transmission lines for microwave electromagnetic waves are rectangular waveguides with so-called partially- and fully-filled waveguides, whose fillable structures are typically comprised of dielectric plates of various shapes and sizes. Artificial materials and structures are used in various waveguide designs, particularly those based on photonic crystals or metamaterials [1–13]. Photonic crystals in the microwave range are based on a section of waveguide whose filling comprises a periodic structure consisting of individual cells made of materials having different refractive indices. This leads to the formation of band gap and allowed photonic bands (frequency ranges) in the transmission spectrum analogous to energy bands in solids [1, 2, 4–9]. Thus, photonic crystals can be used to construct frequency-selective devices, in particular, on the basis of rectangular waveguides [6, 7]. Such devices are capable of isolating microwave radiation both in a specific frequency band (bandpass filter), as well as in the frequency range below (low-pass filter) or above (high-pass filter) a specific cut-off frequency in a given frequency range, and passing it to the output of the device almost without loss.

Adding a single defect to a photonic crystal results in a violation of the periodicity of its structure and the appearance of a resonant transmission peak on its frequency response [10–12]. By controlling the position, shape and size of the defect, resonant (narrow-band) filters, controllable sensors, absorbers, and other useful devices can be created [1–4, 7–10].

This paper presents studies on photonic crystals into which a mobile rotating cylindrical defect has been inserted. In addition, the results of experimental studies into photonic crystal samples fabricated using 3D printing technology are compared with the characteristics of their mathematical models using the specialized *openEMS* software[1].

DESCRIPTION OF THE STRUCTURE

Innovative 3D printing technology is already widely used for the prototyping of products in the microwave range [13–16]. In this paper, fused deposition modeling (FDM), a type of 3D printing technology, is used to fabricate photonic crystals. FDM technology is based around the melting and application of plastic filament to form layers on the surface of previously applied layers to form the structure of a given model. The structure is created by first designing a 3D model, typically in a computer-aided design (CAD) system[2]. The 3D model can be saved in the widely used STL file format as a numerical array. The photonic crystal is designed using the *OpenSCAD* CAD system, which can be used to create complex three-dimensional models with a high degree of parameterization. The design optimization process is greatly simplified by the ability to automatically recalculate the entire geometry by changing one parameter. The resulting photonic crystal comprises a section of a rectangular waveguide having a fully filled cross-section of 23 × 10 mm. Air holes are placed periodically along the waveguide axis. The cross-section of the holes is rectangular. Polyethylene terephthalate glycol (PETG), which has a relative permittivity of $\varepsilon' \approx 2.5$ in the investigated frequency range, is selected as a material for the fabrication of the photonic crystal [15, 16].

A schematic representation of the proposed structure with designations of the main dimensions of the designed

[1] https://www.openems.de. Accessed March 20, 2025.

[2] https://openscad.org. Accessed March 20, 2025.

Table. Main dimensions of the elements of the designed photonic crystal having a rectangular hole shape

Hole shape	Number of holes (i)	Hole period (Λ), mm	Hole size along the waveguide axis (w), mm	Hole spacing (L), mm	Hole size along the wide wall of the waveguide (g), mm
Rectangular	4	24–31	5–9	19–22	18

elements is shown in Fig. 1. The specific values for the dimensions are given in the table. Similar photonic crystal structures have been proposed and analyzed in several works (e.g., [6, 13]).

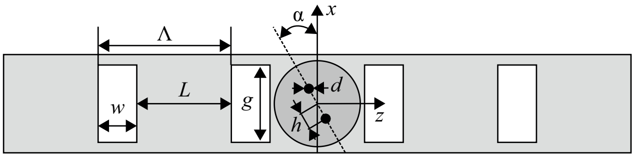


Fig. 1. Schematic representation of the proposed structure for filling a photonic crystal with a mobile rotating defect (viewed from the wide wall side of the waveguide, in the *xz* plane)

A defect located in the center of the designed photonic crystal comprises a movable cylinder rotating around its axis, in which two thin metal rods with a circular cross section and a diameter of $d = 2$ mm are placed symmetrically at a distance of $h = 4$ mm on either side of the axis. The position of the first defect (rotation angle $\alpha = 0°$) corresponds to the position of a pair of defect rods perpendicular to the waveguide axis, along the x-axis. The position of the second defect (rotation angle $\alpha = 90°$) corresponds to the position of a pair of defect rods along the waveguide axis, i.e., along the z-axis.

RESULTS AND DISCUSSION

After determining the optimal structure of the photonic crystal, the stage of numerical modeling of its properties can begin. The *openEMS* software used for this purpose is based on the finite difference time domain method, representing one of the most popular numerical methods in computational electrodynamics [17]. The software supports the import and export of geometric models from various file formats (e.g., PLY, STL), which greatly simplifies the modeling process, especially when CAD is used for design. *openEMS* is integrated with scripting languages such as *MATLAB*[3], *Octave*[4], and *Python*[5] for automating the process of setting model parameters, performing calculations, and processing the obtained data.

Figure 2 depicts 3D printed photonic crystal structures having a cylindrical defect, while the frequency response of the photonic crystals obtained by numerical modeling and experimental studies is shown in Fig. 3. Specific data are given for two defect positions: $\alpha = 0°$ (position *1*) and $\alpha = 90°$ (position *2*). Both numerical simulation and experimental data indicate that a clear resonance peak in transmission can be observed at defect position *2*. At the resonance peak frequency, the change in transmission coefficient (ΔT) exceeds 15 dB when the defect position changes from *1* to *2*.

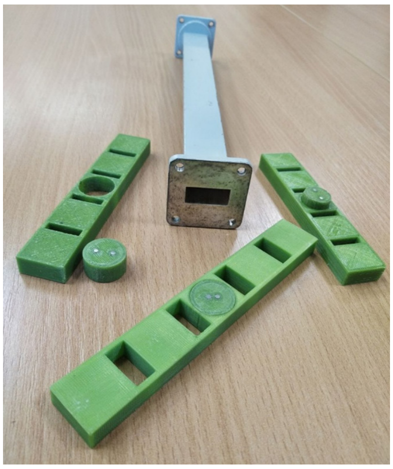


Fig. 2. General view of the manufactured photonic crystals with defect

The dependence of the transmittance peak position on the photonic crystal period is shown in Fig. 4. As the period of the photonic crystal decreases, the transmittance peak is observed to shift towards high frequency. The change in the hole size w has a stronger effect on the position of the transmission peak frequency compared to the change in the hole spacing L.

The dependence of the change in transmission coefficient ΔT on the period of the photonic crystal Λ is shown in Fig. 5. The change in transmission coefficient ΔT reaches 22 dB at the fixed hole size $w = 5$ mm, and the effect of the hole spacing L is minimal. When the hole size w is increased from 3 mm to 9 mm and the hole spacing L is fixed at 22 mm, the change in the transmission coefficient ΔT for the photonic crystal reaches 14 dB (from 16 dB to 30 dB).

[3] https://www.mathworks.com/products/matlab.html. Accessed March 20, 2025.
[4] https://octave.org/. Accessed March 20, 2025.
[5] https://www.python.org/. Accessed March 20, 2025.

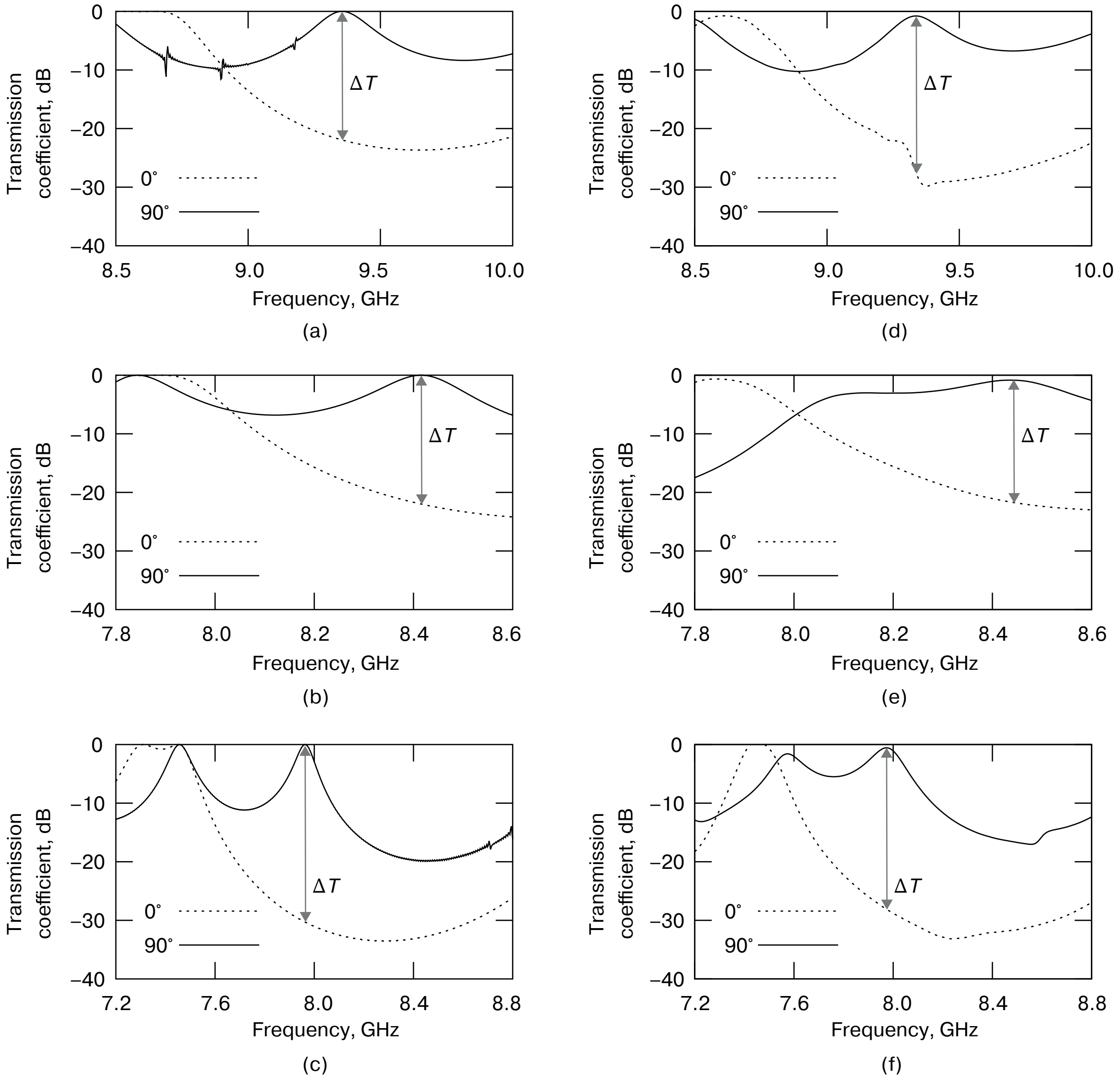


Fig. 3. Frequency response for defect rotation angle α equal to 0° and 90° obtained by numerical modeling (a, b, c) and experimental study (d, e, f) of photonic crystals: L = 19 mm, w = 5 mm (a, d); L = 22 mm, w = 5 mm (b, e); and L = 22 mm, w = 9 mm (c, f)

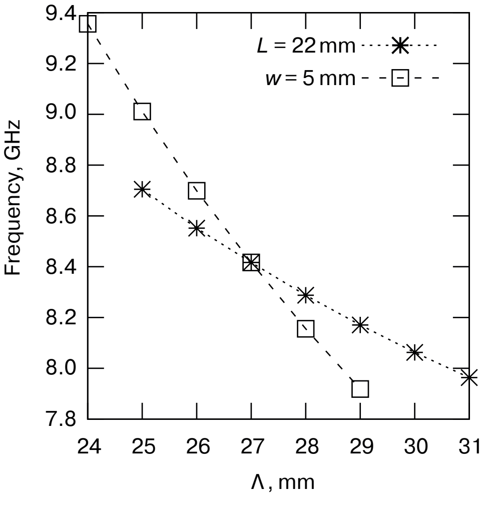


Fig. 4. Dependence of the photonic crystal transmission peak frequency on the period Λ at a fixed hole size w (dotted line) and hole spacing L (dashed line)

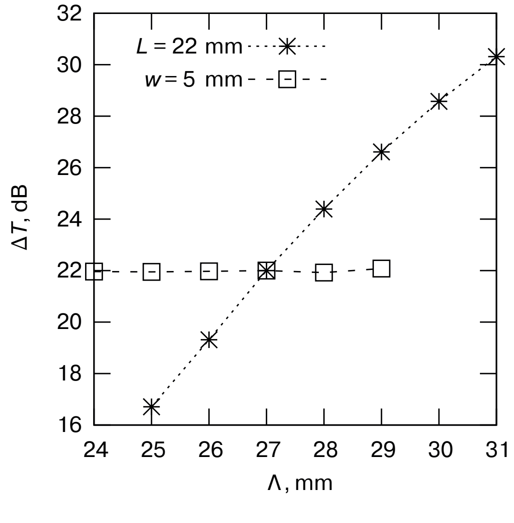


Fig. 5. Dependence of the change in the transmission coefficient (ΔT) on the photonic crystal period Λ at a fixed hole size w (dotted line) and hole spacing L (dashed line)

CONCLUSIONS

The paper demonstrates the possibility of effectively using a mobile cylindrical defect to control the frequency response of a narrow-band filter in a rectangular X-band waveguide having a cross-section of 23 × 10 mm. The defect rotates around its photonic crystal structure axis with respect to which two identical metal pins are symmetrically arranged. Numerical simulation results show that the transmission peak shifts to higher frequencies as the hole period Λ of the photonic crystal decreases. The largest shift in the frequency of the transmission peak with increasing hole period occurs at a fixed hole spacing. At the same time, the minimum change in transmission coefficient is observed for a defect rotation angle of 90° (position *2*). Designed and experimentally developed models of photonic crystal-based narrow-band filters of this structure show a change in the transmission coefficient in the range of 16 dB to 30 dB when the angle of defect rotation is changed from 0° to 90°. The results are promising for use in real technical applications.

Authors' contribution. All authors equally contributed to the research work.

REFERENCES

1. Wan B.F., Xu Y., Zhou Z.W., Zhang D., Zhang H.F. Theoretical investigation of a sensor based on one-dimensional photonic crystals to measure four physical quantities. *IEEE Sens. J*. 2020;21(3):2846–2853. https://doi.org/10.1109/JSEN.2020.3027759
2. Aly A.H., Mohamed D., Mohaseb M.A. Theoretical and simulation study in defective semiconductor layer that incorporated with superconducting-dielectric photonic crystal. *Int. J. Modern Phys. B*. 2019;33(32):1950397. https://doi.org/10.1142/S0217979219503971
3. Buchnev I.Yu., Osipov O.V. Investigation of the electromagnetic properties of a transverse insert based on a planar layer of a chiral metamaterial in a rectangular waveguide. *Fizika volnovykh protsessov i radiotekhnicheskie sistemy = Physics of Wave Processes and Radio Systems*. 2023;26(1):93–105 (in Russ.). https://doi.org/10.18469/1810-3189.2023.26.1.93-105
4. Ghasemi F., Aliasghary M., Razi S. Magneto-sensitive photonic crystal optical filter with tunable response in 12–19 GHz; cross over from design to prediction of performance using machine learning. *Phys. Lett. A*. 2021;401:127328. https://doi.org/10.1016/j.physleta.2021.127328
5. Zhao L., Li Y., Chen Z.M., Liang Z.H., Wang J., Shen X., Zhang Q. A Band-Pass Filter Based on Half-Mode Substrate Integrated Waveguide and Spoof Surface Plasmon Polaritons. *Sci Rep*. 2019;9(1):13429. https://doi.org/10.1038/s41598-019-50056-9
6. Mikhailov A.I., Ryabov E.A., Sergeev S.A. Evaluation of the possibilities of 3D-printing for the making of waveguide photonic crystals. *Fizika volnovykh protsessov i radiotekhnicheskie sistemy = Physics of Wave Processes and Radio Systems*. 2022;25(3):29–35 (in Russ.). https://doi.org/10.18469/1810-3189.2022.25.3.29-35
7. Ishchenko E.A., Pasternak Yu.G., Pendyurin V.A., Fyedorov S.M., Chernoivanenko I.A. Active rectangular waveguide bandpass filter based on the metamaterial. *Vestnik Voronezhskogo gosudarstvennogo tekhnicheskogo universiteta = Bulletin of* the *Voronezh State Technical University*. 2022;18(3):57–60 (in Russ.). https://doi.org/10.36622/VSTU.2022.18.3.007
8. Liu L., Ye M., Yu Z., Xue W. Notch microwave photonic filter with narrow bandwidth and ultra-high all-optical tuning efficiency based on a silicon nanobeam cavity. *J. Lightwave Technol*. 2023;41(15):5051–5058. https://doi.org/10.1109/JLT.2023.3248611
9. Yu B., Yang J., Song Y., Wang Z., Zhang T., Yan B., Xu R. Terahertz Metamaterial Waveguide with I-Shaped Resonators for Phase and Absorption Modulation. *Photonics*. 2023;10(7):816. https://doi.org/10.3390/photonics10070816
10. Skripal A.V., Ponomarev D.V., Sharonov V.E. Resonance characteristics of microwave photonic crystals with inclusions in the form of conducting nanolayers. *Technical Physics Letters*. 2023;10:23–26.
    [Original Russian Text: Skripal A.V., Ponomarev D.V., Sharonov V.E. Resonance characteristics of microwave photonic crystals with inclusions in the form of conducting nanolayers. *Pis'ma v Zhurnal tekhnicheskoi fiziki*. 2023;49(19):27–30 (in Russ.). https://doi.org/10.61011/PJTF.2023.19.56269.19645]
11. Kumar N., Pandey G.N., Dhayal S., Dhayal S.S. Microwave Propagation Characteristics in Magnetized-Cold-Plasma-Based Binary Photonic Crystal with Defect of MCP Layer. *Macromol. Symp*. 2023;407(1):2100515. https://doi.org/10.1002/masy.202100515
12. Usanov D.A., Skripal' A.V., Posadskii V.N., et al. Defect Mode in Microwave Waveguide Bragg Structures with Metal Pins. *Tech. Phys*. 2019;64(10):1523–1526. https://doi.org/10.1134/S1063784219100232
    [Original Russian Text: Usanov D.A., Skripal' A.V., Posadskii V.N., Tyazhlov V.S., Baikin A.V. Defect Mode in Microwave Waveguide Bragg Structures with Metal Pins. *Zhurnal tekhnicheskoi fiziki*. 2019;89(10):1606–1610 (in Russ.). https://doi.org/10.21883/JTF.2019.10.48180.6-19]
13. Khairushev I.V., Ryabov E.A., Sergeev S.A Theoretical and experimental studies of photonic crystals manufactured by 3D printing technology in the X-band. *Microwave Electronics and Microelectronic*s. 2022;1:546–549 (in Russ.). Available from URL: https://mwelectronics.etu.ru/assets/files/2022/546-549.pdf
14. Pei Z., Xu Y., Wei F., Liu T., Su D. Electromagnetic property of a novel gradient honeycomb composite fabricated by 3D forming. *J. Magn. Magn. Mater*. 2020;493:165742. https://doi.org/10.1016/j.jmmm.2019.165742

15. Andreev A.A., Ryabov E.A., Khairushev I.V., Sergeeva B.V., Sergeev S.A. The influence of temperature on the dielectric constant of plastics in the microwave range. *Microwave Electronics and Microelectronics*. 2023;1:388–392 (in Russ.). Available from URL: https://mwelectronics.etu.ru/assets/files/2023/novoe/388-392.pdf
16. Zechmeister J., Lacik J. Complex Relative Permittivity Measurement of Selected 3D-Printed Materials up to 10 GHz. In: *2019 Conference on Microwave Techniques* (*COMITE*). IEEE; 2019. P. 1–4. https://doi.org/10.1109/COMITE.2019.8733590
17. Liebig T., Rennings A., Held S., Erni D. OpenEMS – a free and open source equivalent-circuit (EC) FDTD simulation platform supporting cylindrical coordinates suitable for the analysis of traveling wave MRI applications. *Int. J. Numerical Modelling: Electronic Networks, Devices and Fields*. 2013;26(6):680–696. https://doi.org/10.1002/jnm.1875

СПИСОК ЛИТЕРАТУРЫ

1. Wan B.F., Xu Y., Zhou Z.W., Zhang D., Zhang H.F. Theoretical investigation of a sensor based on one-dimensional photonic crystals to measure four physical quantities. *IEEE Sens. J*. 2020;21(3):2846–2853. https://doi.org/10.1109/JSEN.2020.3027759
2. Aly A.H., Mohamed D., Mohaseb M.A. Theoretical and simulation study in defective semiconductor layer that incorporated with superconducting-dielectric photonic crystal. *Int. J. Modern Phys. B*. 2019;33(32):1950397. https://doi.org/10.1142/S0217979219503971
3. Бучнев И.Ю., Осипов О.В. Исследование электромагнитных свойств поперечной вставки на основе планарного слоя кирального метаматериала в прямоугольном волноводе. *Физика волновых процессов и радиотехнические системы*. 2023;26(1):93–105. https://doi.org/10.18469/1810-3189.2023.26.1.93-105
4. Ghasemi F., Aliasghary M., Razi S. Magneto-sensitive photonic crystal optical filter with tunable response in 12–19 GHz; cross over from design to prediction of performance using machine learning. *Phys. Lett. A*. 2021;401:127328. https://doi.org/10.1016/j.physleta.2021.127328
5. Zhao L., Li Y., Chen Z.M., Liang Z.H., Wang J., Shen X., Zhang Q. A Band-Pass Filter Based on Half-Mode Substrate Integrated Waveguide and Spoof Surface Plasmon Polaritons. *Sci Rep*. 2019;9(1):13429. https://doi.org/10.1038/s41598-019-50056-9
6. Михайлов А.И., Рябов Е.А., Сергеев С.А. Оценка возможностей 3D-печати для изготовления волноводных фотонных кристаллов. *Физика волновых процессов и радиотехнические системы*. 2022;25(3):29–35. https://doi.org/10.18469/1810-3189.2022.25.3.29-35
7. Ищенко Е.А., Пастернак Ю.Г., Пендюрин В.А., Фёдоров С.М., Черноиваненко И.А. Активный волноводный полосовой фильтр на основе метаматериала. *Вестник Воронежского государственного технического университета*. 2022;18(3):57–60. https://doi.org/10.36622/VSTU.2022.18.3.007
8. Liu L., Ye M., Yu Z., Xue W. Notch microwave photonic filter with narrow bandwidth and ultra-high all-optical tuning efficiency based on a silicon nanobeam cavity. *J. Lightwave Technol*. 2023;41(15):5051–5058. https://doi.org/10.1109/JLT.2023.3248611
9. Yu B., Yang J., Song Y., Wang Z., Zhang T., Yan B., Xu R. Terahertz Metamaterial Waveguide with I-Shaped Resonators for Phase and Absorption Modulation. *Photonics*. 2023;10(7):816. https://doi.org/10.3390/photonics10070816
10. Скрипаль А.В., Пономарев Д.В., Шаронов В.Е. Резонансные характеристики сверхвысокочастотных фотонных кристаллов с включениями в виде проводящих нанослоев. *Письма в Журнал технической физики* (*Письма в ЖТФ*). 2023;49(19):27–30. https://doi.org/10.61011/PJTF.2023.19.56269.19645
11. Kumar N., Pandey G.N., Dhayal S., Dhayal S.S. Microwave Propagation Characteristics in Magnetized-Cold-Plasma-Based Binary Photonic Crystal with Defect of MCP Layer. *Macromol. Symp*. 2023;407(1):2100515. https://doi.org/10.1002/masy.202100515
12. Усанов Д.А., Скрипаль А.В., Посадский В.Н., Тяжлов В.С., Байкин А.В. Дефектная мода в СВЧ волноводных брэгговских структурах с металлическими штырями. *Журнал технической физики*. 2019;89(10):1606–1610. https://doi.org/10.21883/JTF.2019.10.48180.6-19
13. Хайрушев И.В., Рябов Е.А., Сергеев С.А. Теоретические и экспериментальные исследования фотонных кристаллов, изготовленных технологией 3D-печати, в X-диапазоне. *Электроника и микроэлектроника СВЧ*. 2022;1:546–549. URL: https://mwelectronics.etu.ru/assets/files/2022/546-549.pdf
14. Pei Z., Xu Y., Wei F., Liu T., Su D. Electromagnetic property of a novel gradient honeycomb composite fabricated by 3D forming. *J. Magn. Magn. Mater*. 2020;493:165742. https://doi.org/10.1016/j.jmmm.2019.165742
15. Андреев А.А., Рябов Е.А., Хайрушев И.В., Сергеева Б.В., Сергеев С.А. Влияние температуры на диэлектрическую проницаемость пластиков в СВЧ диапазоне. *Электроника и микроэлектроника СВЧ*. 2023;1:388–392. URL: https://mwelectronics.etu.ru/assets/files/2023/novoe/388-392.pdf
16. Zechmeister J., Lacik J. Complex Relative Permittivity Measurement of Selected 3D-Printed Materials up to 10 GHz. In: *2019 Conference on Microwave Techniques* (*COMITE*). IEEE; 2019. P. 1–4. https://doi.org/10.1109/COMITE.2019.8733590
17. Liebig T., Rennings A., Held S., Erni D. OpenEMS – a free and open source equivalent-circuit (EC) FDTD simulation platform supporting cylindrical coordinates suitable for the analysis of traveling wave MRI applications. *Int. J. Numerical Modelling: Electronic Networks, Devices and Fields*. 2013;26(6):680–696. https://doi.org/10.1002/jnm.1875

About the Authors

Evgeny A. Ryabov, Assistant, Department of Solid State Physics, Institute of Physics, Saratov National Research State University (83, Astrakhanskaya ul., Saratov, 410012 Russia). E-mail: k1u2r3ka@mail.ru. RSCI SPIN-code 9110-4151, https://orcid.org/0000-0003-4777-7346

Anton A. Andreev, Engineer, Semiconductor Technology Educational Laboratory, Institute of Physics, Saratov National Research State University (83, Astrakhanskaya ul., Saratov, 410012 Russia). E-mail: andreev25304@mail.ru. RSCI SPIN-code 6173-0839, https://orcid.org/0009-0003-3212-6484

Sergey A. Sergeev, Cand. Sci. (Phys.-Math.), Associate Professor, Department of Solid State Physics, Institute of Physics, Saratov National Research State University (83, Astrakhanskaya ul., Saratov, 410012 Russia). E-mail: ssergeev@bk.ru. RSCI SPIN-code 6883-7787, https://orcid.org/0000-0002-4442-6797

Alexander I. Mikhailov, Dr. Sci. (Phys.-Math.), Professor, Department of State Physics, Institute of Physics, Saratov National Research State University (83, Astrakhanskaya ul., Saratov, 410012 Russia). E-mail: mikhailovai13@mail.ru. RSCI SPIN-code 2491-0488, https://orcid.org/0000-0002-4158-9195

Об авторах

Рябов Евгений Александрович, ассистент, кафедра физики твердого тела, Институт физики, ФГБОУ ВО «Саратовский национальный исследовательский государственный университет имени Н.Г. Чернышевского» (410012, Россия, Саратов, ул. Астраханская, д. 83). E-mail: k1u2r3ka@mail.ru. SPIN-код РИНЦ 9110-4151, https://orcid.org/0000-0003-4777-7346

Андреев Антон Андреевич, инженер, учебная лаборатория по полупроводниковой электронике, Институт физики, ФГБОУ ВО «Саратовский национальный исследовательский государственный университет имени Н.Г. Чернышевского» (410012, Россия, Саратов, ул. Астраханская, д. 83). E-mail: andreev25304@mail.ru. SPIN-код РИНЦ 6173-0839, https://orcid.org/0009-0003-3212-6484

Сергеев Сергей Алексеевич, к.ф.-м.н., доцент, кафедра физики твердого тела, Институт физики, ФГБОУ ВО «Саратовский национальный исследовательский государственный университет имени Н.Г. Чернышевского» (410012, Россия, Саратов, ул. Астраханская, д. 83). E-mail: ssergeev@bk.ru. SPIN-код РИНЦ 6883-7787, https://orcid.org/0000-0002-4442-6797

Михайлов Александр Иванович, д.ф.-м.н., профессор, кафедра физики твердого тела, Институт физики, ФГБОУ ВО «Саратовский национальный исследовательский государственный университет имени Н.Г. Чернышевского» (410012, Россия, Саратов, ул. Астраханская, д. 83). E-mail: mikhailovai13@mail.ru. SPIN-код РИНЦ 2491-0488, https://orcid.org/0000-0002-4158-9195

Translated from Russian into English by K. Nazarov
Edited for English language and spelling by Thomas A. Beavitt

Modern radio engineering and telecommunication systems

Современные радиотехнические и телекоммуникационные системы

UDC 621.372.8
https://doi.org/10.32362/2500-316X-2025-13-3-92-102
EDN OEBXOF

RESEARCH ARTICLE

Development of a microwave low-pass filter based on a microstrip line projection model

Alexey D. Yarlykov [@],
Oleg A. Demin

MIREA – Russian Technological University, Moscow, 119454 Russia
[@] *Corresponding author, e-mail: yarlykov@mirea.ru*

• **Submitted:** 18.09.2024 • **Revised:** 03.02.2025 • **Accepted:** 24.03.2025

Abstract
Objectives. Sections of microstrip lines having finite length are widely used to develop integrated circuits and microwave devices for various purposes, such as power dividers, directional couplers, attenuators, and filters. In particular, low-pass filters in the microwave range are comprised of a cascade connection of regular sections of microstrip lines having various geometric parameters. However, modern approaches to calculating microwave filters using commercial software require large computational and time-consuming resources, especially when carrying out electrodynamic analysis of microstrip lines. The work set out to develop an algorithm and a method for calculating filters using a projection approach to the electrodynamic analysis of microstrip lines that reduces the time required to calculate characteristics of microwave filters while maintaining high accuracy of the obtained results.
Methods. The proposed projection approach to the electrodynamic analysis of a microstrip line can be used to rapidly and accurately calculate the main electrodynamic parameters of retardation coefficient and wave impedance across a wide range of changes in the geometrical parameters of the line, as well as its dielectric constant and frequency.
Results. Formulas obtained on the basis of analytical expressions for calculating the electrodynamic parameters of a microstrip line are used to describe the nature of changes in the elements of the scattering matrix of multistage low-pass filters in a given frequency band. A developed computer program was used to calculate the values of the elements of the low-pass filter scattering matrix across a wide range of substrate dielectric constant and frequency parameters. The obtained results were compared with the characteristics of filters calculated using commercial software.
Conclusions. The proposed approach to calculating the electrodynamic parameters of microstrip lines and consequent elements of the scattering matrix of multistage low-pass filters can significantly reduce the calculation time while achieving a sufficiently high accuracy of the obtained results to significantly reduce labor costs when calculating microwave filters in engineering practice.

Keywords: microstrip line, projection approach, low-pass filter, retardation coefficient, wave impedance, scattering matrix, reflection coefficient, transmission coefficient

© A.D. Yarlykov, O.A. Demin, 2025

For citation: Yarlykov A.D., Demin O.A. Development of a microwave low-pass filter based on a microstrip line projection model. *Russian Technological Journal.* 2025;13(3):92–102. https://doi.org/10.32362/2500-316X-2025-13-3-92-102, https://www.elibrary.ru/OEBXOF

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Разработка сверхвысокочастотного фильтра нижних частот на основе проекционной модели микрополосковой линии

А.Д. Ярлыков [@],
О.А. Демин

МИРЭА – Российский технологический университет, Москва, 119454 Россия
[@] Автор для переписки, e-mail: yarlykov@mirea.ru

• **Поступила:** 18.09.2024 • **Доработана:** 03.02.2025 • **Принята к опубликованию:** 24.03.2025

Резюме

Цели. Отрезки микрополосковых линий конечной длины применяются для разработки интегральных схем и устройств сверхвысоких частот (СВЧ) различного назначения, таких как делители мощности, направленные ответвители, аттенюаторы и фильтры, имеющих, в большинстве случаев, сложную топологическую структуру. В частности, фильтры нижних частот (ФНЧ) СВЧ-диапазона представляют собой ступенчатое соединение регулярных отрезков микрополосковых линий с различными геометрическими параметрами. Однако современные подходы к расчету фильтров СВЧ при помощи коммерческих программ требуют больших вычислительных и временных затрат, связанных, в частности, с предложенными подходами к электродинамическому анализу микрополосковых линий. Целью статьи является разработка алгоритма и методики расчета фильтров с использованием проекционного подхода к электродинамическому анализу микрополосковых линий, позволяющих сократить время расчета характеристик фильтров СВЧ при сохранении высокой точности полученных результатов.

Методы. Предложен проекционный подход к проведению электродинамического анализа микрополосковой линии, позволяющий быстро и с высокой точностью проводить расчет ее основных электродинамических параметров – коэффициента замедления и волнового сопротивления в широком диапазоне изменения геометрических параметров линии, ее диэлектрической проницаемости и частоты.

Результаты. На базе аналитических выражений для расчета электродинамических параметров микрополосковой линии получены формулы для описания характера изменений элементов матрицы рассеяния многокаскадных ФНЧ в заданной полосе частот. Разработана компьютерная программа, позволяющая рассчитывать значения элементов матрицы рассеяния ФНЧ в широком диапазоне диэлектрической проницаемости подложки и частоты. Проведено сравнение полученных результатов с характеристиками фильтров, рассчитанных при помощи коммерческих программ.

Выводы. Предложенный подход к расчету электродинамических параметров микрополосковых линий и, как следствие, элементов матрицы рассеяния многокаскадных ФНЧ позволяет значительно сократить время расчетов при достижении достаточно высокой точности полученных результатов, что значительно снижает трудозатраты при проектировании фильтров СВЧ в инженерной практике.

Ключевые слова: микрополосковая линия, проекционный подход, фильтр нижних частот, коэффициент замедления, волновое сопротивление, матрица рассеяния, коэффициент отражения, коэффициент передачи

Для цитирования: Ярлыков А.Д., Демин О.А. Разработка сверхвысокочастотного фильтра нижних частот на основе проекционной модели микрополосковой линии. *Russian Technological Journal*. 2025;13(3):92–102. https://doi.org/10.32362/2500-316X-2025-13-3-92-102, https://www.elibrary.ru/OEBXOF

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

Today, the vast majority of microwave devices and modules are structurally based on microstrip transmission lines (MTL). This is due to their small mass and size parameters, as well as the ease of transition to their topology from elements offering concentrated parameters during low-frequency prototyping of microwave devices [1]. For example, the simplest microwave low-pass filter (LPF) topology comprises a cascade of regular sections of finite-length MTLs having different strip conductor widths [2]. The geometric parameters of the MTL, as well as the dielectric constant and frequency of its substrate, determine the value of the main electrodynamic parameters of the line, namely the retardation coefficient and wave impedance, which play a key role in calculating the characteristics of microwave filters [3]. However, contemporary approaches to their calculation using a number of commercially available software products imply rather high computational and time costs both in the calculation of the basic electrodynamic parameters of MTLs [4] and in the design of microwave filters in general [5]. Considering this, the task of applying the MTL projection model detailed in [6] is relevant. In this model, an open MTL is simulated as a shield over a wide range of geometrical parameters, as well as substrate permittivity and frequency. Furthermore, [7] defines the minimum shield size that allows an open MTL to be modelled with a given accuracy, while [8] presents methods to improve the efficiency of the proposed MTL model. This paper describes the application of the MTL model proposed in [9] to the calculation of the LPF microwave scattering matrix. The presented model demonstrates high accuracy of results along with a significant reduction in the time requirement.

1. DESIGN METHODOLOGY FOR MICROWAVE LPFS

The basic methodology of microwave filter design is given in [10–12]. According to the classical approach, the first stage of microwave filter design is the calculation of its low-frequency prototype on concentrated elements. It starts with the determination of the number of sections (links) in the calculated filter. This number is determined based on the required type of filter approximation and the amount of barrier band attenuation. In this way, for the Butterworth filter, the number of sections can be determined by the following formula:

$$N = \frac{\lg\left(10^{\left[L(\omega)/10\right]} - 1\right)}{2\lg\left(\omega/\omega_{\mathrm{CO}}\right)}, \tag{1}$$

and for the Chebyshev filter:

$$N = \frac{\operatorname{arch}\left\{\left(10^{\left[L(\omega)/10\right]} - 1\right)\Big/\left(10^{\left[G_{\mathrm{T}}/10\right]} - 1\right)\right\}^{1/2}}{\operatorname{arch}\left(\omega/\omega_{\mathrm{CO}}\right)}, \tag{2}$$

where $L(\omega)$ is the value of attenuation at frequency ω in the cut-off band; ω_{CO} is the filter cut-off frequency; G_{T} is the pulse amplitude (Throb) in the passband (in decibels).

Having determined the number of links in the filter, its equivalent low-frequency circuit can be constructed in which each filter section is represented as a concentrated element (inductance or capacitance) according to Table 5.2 from [12]. The circuit shown in Fig. 1 is an example of such an equivalent circuit for a five-link LPF.

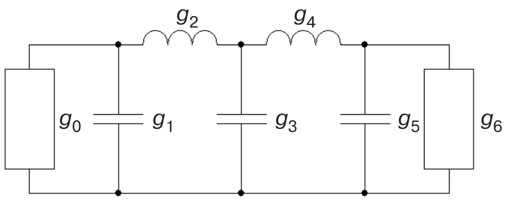


Fig. 1. Equivalent circuit of the low-pass prototype of the five-section LPF. g_i stands for normalized parameters of the equivalent circuit

The normalized parameters of the equivalent circuit (g-parameters) are determined by the type of filter approximation and the total number of sections N. For the Butterworth filter, the g-parameters are determined by the following formulae:

$$g_0 = g_{N+1} = 1,$$
$$g_k = 2\sin\left[\frac{(2k-1)\pi}{2N}\right],\ k=\overline{1,N}, \tag{3}$$

and for the Chebyshev filter by

$$\begin{gathered}
g_0 = 1,\\
g_1 = \frac{2a_1}{\psi},\\
g_k = \frac{4a_{k-1}a_k}{b_{k-1}g_{k-1}},\ k=\overline{2,N},\\
g_{N+1} = \begin{cases} 1 \text{ at uneven } N,\\ \operatorname{cth}^2(\beta/4) \text{ at even } N,\end{cases}\\
\beta = \ln\left[\operatorname{cth}\left(G_\mathrm{T}/17.37\right)\right],\\
\psi = \operatorname{sh}\left(\frac{\beta}{2N}\right),\\
a_k = \sin\left[\frac{(2k-1)\pi}{2N}\right],\\
b_k = \psi^2 + \sin^2\left(\frac{\pi k}{N}\right).
\end{gathered} \tag{4}$$

Once the g-parameters have been calculated, they should be denormalized to determine the absolute values of the capacitances C_k and inductances L_k in the equivalent circuit, as well as the generator and load resistances R_k, thus determining the wave impedance of the supply line. Denormalization is carried out according to the following rule:

$$R_k = R_\mathrm{LD}g_k,\ C_k = \frac{g_k}{R_\mathrm{LD}\omega_\mathrm{CO}},\ L_k = \frac{g_k R_\mathrm{LD}}{\omega_\mathrm{CO}}, \tag{5}$$

where R_LD is the load resistance equal to the wave impedance of the supply line.

Having calculated the parameters of the equivalent circuit, a transition to the topology of the filter on distributed elements should be carried out according to Table 5.3 from [10]. An example of the topology of a five-section LPF on MTL is shown in Fig. 2. As can be seen from the figure, this consists of a linear strip conductor with a varying strip width W along its length. The section l_1 of the MTL has a large wave impedance with respect to the wave impedance Z of the supply line, while the section l_2 has a smaller wave impedance. If $l_1 < \frac{\lambda_\mathrm{CO}}{8}$ and $l_2 < \frac{\lambda_\mathrm{CO}}{8}$ $\left(\lambda_\mathrm{CO} = \frac{6\pi\cdot 10^8}{\omega_\mathrm{CO}}\right)$, then section l_1 has an inductive impedance and l_2 has a capacitive impedance, where λ_CO is the wavelength corresponding to the cut-off frequency. Therefore, the above conditions should be checked when selecting the wave impedance and calculating the length of the sections.

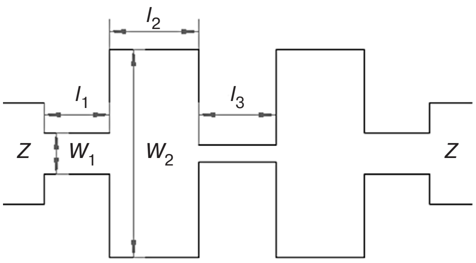


Fig. 2. Topology of a five-section microwave LPF on an MTL. W_1, W_2 are the widths of the strip conductors

In order to ensure a single-wave mode in the line (no transverse resonance), the width of the strip conductor should not exceed $\frac{\lambda_\mathrm{CO}}{4}$.

To ensure a jump in resistance at the transition from inductive to capacitive element and vice versa, the ratio of wave impedances for these elements should not be less than 3 times. The width of the strip conductors, the dielectric permittivity of the substrate, and its height are determined on the basis of the required wave impedances and taking into account the conditions described above.

Once the wave impedances and the widths of the strip conductors have been selected, the lengths of all the line segments in the filter should be determined. The length of the segment that implements the inductance is determined by the following formula:

$$l_L = \frac{3\cdot 10^8}{\omega_\mathrm{CO}\sqrt{\varepsilon}}\arcsin\left(\frac{\omega_\mathrm{CO}L}{Z_L}\right), \tag{6}$$

while the length of the segment that implements the capacitance is given by

$$l_C = \frac{3\cdot 10^8}{\omega_\mathrm{CO}\sqrt{\varepsilon}}\arcsin\left(\omega_\mathrm{CO}CZ_C\right), \tag{7}$$

where ε is the dielectric permittivity of the substrate; L and C are the inductance and capacitance values calculated at the low-frequency prototyping stage, respectively; and Z_L and Z_C are the wave resistances of the inductive and capacitive MTL segments, respectively.

The final stage in the design of the filter topology is the correction of the lengths of its capacitive and inductive line segments, taking into account the influence of terminal capacitances and inductances. Their values are subtracted from the initial values of inductance and capacitance for each section of the filter. Taking into account the obtained values, the corrected length of the line segments is calculated again using Eqs. (6) and (7).

The most labor-intensive part of the whole design process consists in the determination of the geometrical parameters of the strip conductor segments based on the required ratio of the line wave impedances, which can be determined using special graphs, for example, in [13]. However, it should be noted that all elements of the matrix are frequency-dependent functions when calculating the scattering matrix of the microwave filter. This requires the calculation of wave impedances and propagation constants of MTL segments as dispersion properties, which leads to significant computation time when modeling filters using commercial software.

2. PROJECTION APPROACH FOR THE CALCULATION OF MTL ELECTRODYNAMIC PARAMETERS

The projection approach considered in [14] for reducing the computational and time costs represents the surface current density on the strip conductor as a system of basic functions in the form of Chebyshev polynomials that take into account the field characteristics at the edges of the strip conductor. By decomposing the longitudinal component of the surface current density by the Chebyshev basis of only one basis function, the dispersion equation used to determine the MTL retardation coefficient n_0 can be obtained as follows:

$$\sum_{m=1}^{\infty}\left[\frac{1}{\chi_m^2}\left(n_0^2 G_m^{\mathrm{E}}+\alpha_m^2 G_m^{\mathrm{M}}\right)\right]J_0^2(m\alpha)\sin^2(m\beta)=0, \quad (8)$$

where $G_m^{\mathrm{E}}=\left(\frac{\varepsilon}{\beta_{m1}}\mathrm{ctg}\left(k_0\beta_{m1}h\right)+\frac{1}{\beta_{m2}}\mathrm{ctg}\left[k_0\beta_{m2}(b-h)\right]\right)^{-1}$, $G_m^{\mathrm{M}}=\left(\beta_{m1}\mathrm{ctg}\left(k_0\beta_{m1}h\right)+\beta_{m2}\mathrm{ctg}\left[k_0\beta_{m2}(b-h)\right]\right)^{-1}$ are the functions obtained by solving the electric (E) and magnetic (M) eigenwave problems, respectively; $J_0(m\alpha)$ is the Bessel function; m is an integer that determines the field structure in MTL; h is the substrate height; $\beta_{m1}=\sqrt{\varepsilon-\chi_m^2}$; $\beta_{m2}=\sqrt{1-\chi_m^2}$; $\chi_m^2=\alpha_m^2+\mathrm{G}^2$; $\alpha_m=\frac{\pi}{k_0 a}m$; $\alpha=\frac{\pi}{2}\cdot\frac{W}{a}$; $\beta=\alpha\left(1+\frac{S}{W}\right)$; $k_0 = 2\pi f/c$ is the wave number; f is the frequency; c is the speed of light in a vacuum; W is the strip width; a, b are the shield dimensions; and S is the distance from the strip edge to the shield wall.

The wave impedance Z_0 is determined by the power carried through the line cross section and the current in the strip conductor, as follows:

$$Z_0=\frac{240\pi}{k_0 a}n\times$$
$$\times\sum_{m=1}^{\infty}\left[-\left(n_0^2\left(G_m^{\mathrm{E}}\right)'+\alpha_m^2\left(G_m^{\mathrm{M}}\right)'\right)+\frac{\alpha_m^2}{\chi_m^2}\left(G_m^{\mathrm{M}}-G_m^{\mathrm{E}}\right)\right]\times \quad (9)$$
$$\times\frac{1}{\chi_m^2}J_0^2(m\alpha)\sin^2(m\beta),$$

where $\left(G_m^{\mathrm{E,M}}\right)'$ is the derivative of function $G_m^{\mathrm{E,M}}$ with respect to n_0^2.

In [9], the problem of calculating the retardation coefficient and the wave impedance of the shielded MTL using Eqs. (8) and (9) is considered. The slow convergence of the series included in these expressions results in a considerable time needed to calculate the parameters to ensure the convergence of the series. Thus, simple formulae for calculating the main electrodynamic parameters of the line in the quasi-static approximation are proposed along with the limits of their applicability. After determining the dependence of retardation coefficient and wave impedance on frequency based on the obtained expressions, the following simple formulas can be used for calculating dispersion characteristics of shielded MTL:

$$n(f)\approx\xi_n n_0(f),\ Z(f)\approx\xi_Z Z_0(f), \quad (10)$$

where $Z_0(f)$ is the wave impedance calculated in the "first approximation" at a given frequency f; ξ_n, ξ_Z are coefficients that depend on the width of the strip conductor and are determined by the following approximation formulae:

$$\xi_n\approx\begin{cases}1,\ \text{at}\ W/h<2,\\ 1+3.6\cdot10^{-3}\left(W/h-2\right),\ \text{at}\ 2\le W/h\le10,\end{cases}$$
$$\xi_Z\approx\begin{cases}1,\ \text{at}\ W/h<2,\\ 1-8.725\cdot10^{-3}\left(W/h-2\right),\ \text{at}\ 2\le W/h\le10.\end{cases} \quad (11)$$

Given the proposed formulae, the table shows the values of the propagation constant and wave impedance of MTL over a wide range of variations of the strip conductor width, the dielectric constant of the substrate, and the frequency. The first lines of each cell show the values of the propagation constant G = nk_0 and the wave impedance Z for a line with a quartz substrate (ε = 3.8). The second lines show the values for a line with a polycore substrate (ε = 9.6), while the third lines show the values for a line with an arsenide-gallium substrate (ε = 13.3). Here, the width of the strip conductor and the frequency are normalized to the substrate height h to unify the values of the electrodynamic parameters of the line. These tables allow the design engineer to determine the material and substrate height, as well as the geometric parameters of the microwave filter topology, on the basis of the specified cut-off frequency. The numerical values of the propagation constant and the retardation coefficient given for the selected geometrical and physical parameters of the filter are also necessary for the calculation of the scattering matrix elements.

Table. MTL propagation constant and wave impedance

fh, GHz · mm		0.1	1	3	5	7	10	15
W/h = 0.1	G	0.19 0.29 0.34	1.93 2.90 3.37	5.79 8.72 10.16	9.67 14.60 17.04	13.57 20.55 24.03	19.45 29.63 34.72	29.38 45.15 53.13
	Z	163.70 109.01 93.66	163.71 108.99 93.63	163.85 109.05 93.68	164.16 109.34 94.01	164.67 109.92 94.70	165.81 111.45 96.56	168.81 116.12 102.46
W/h = 0.5	G	0.20 0.30 0.35	1.97 2.98 3.48	5.92 8.99 10.50	9.90 15.10 17.67	13.90 21.31 25.01	19.97 30.85 36.31	30.26 47.28 55.92
	Z	101.64 67.23 57.69	101.65 67.21 57.66	101.75 67.23 57.67	102.00 67.44 57.91	102.40 67.90 58.44	103.29 69.06 59.83	105.57 72.36 63.82
W/h = 1	G	0.20 0.30 0.36	2.01 3.05 3.56	6.03 9.22 10.79	10.09 15.51 18.20	14.18 21.94 25.81	20.40 31.83 37.58	30.97 48.90 57.96
	Z	75.74 49.82 42.70	75.74 49.80 42.67	75.82 49.80 42.66	76.03 49.97 42.86	76.36 50.35 43.31	77.09 51.31 44.42	78.91 53.82 47.31
W/h = 1.5	G	0.20 0.31 0.36	2.03 3.11 3.63	6.12 9.41 11.02	10.24 15.85 18.63	14.41 22.44 26.44	20.74 32.59 38.53	31.51 50.09 59.40
	Z	61.45 40.24 34.46	61.45 40.21 34.42	61.52 40.21 34.41	61.69 40.37 34.60	61.98 40.70 34.99	62.61 41.51 35.93	64.13 43.51 38.16
W/h = 2	G	0.21 0.32 0.37	2.06 3.16 3.69	6.19 9.56 11.22	10.37 16.12 18.98	14.59 22.85 26.96	21.02 33.20 39.28	31.95 50.98 60.46
	Z	52.01 33.93 29.03	52.01 33.91 29.00	52.06 33.91 28.99	52.22 34.07 29.17	52.48 34.40 29.54	53.05 35.15 30.36	54.36 36.89 32.16
W/h = 3	G	0.21 0.32 0.38	2.09 3.23 3.79	6.31 9.80 11.52	10.57 16.55 19.51	14.88 23.47 27.73	21.45 34.08 40.35	32.61 52.22 61.89
	Z	40.06 25.99 22.21	40.05 25.97 22.18	40.10 25.98 22.19	40.23 26.15 22.37	40.46 26.46 22.69	40.93 27.09 23.34	41.97 28.41 24.60
W/h = 4	G	0.21 0.33 0.38	2.12 3.29 3.85	6.39 9.99 11.75	10.72 16.87 19.90	15.10 23.91 28.26	21.76 34.69 41.06	33.06 53.01 62.78
	Z	32.71 21.14 18.05	32.71 21.12 18.03	32.75 21.15 18.05	32.87 21.32 18.22	33.08 21.60 18.51	33.49 22.14 19.04	34.35 23.17 19.99
W/h = 5	G	0.21 0.33 0.39	2.14 3.33 3.91	6.46 10.13 11.92	10.83 17.11 20.19	15.26 24.24 28.65	22.00 35.12 41.56	33.40 53.55 63.36
	Z	27.70 17.85 15.24	27.70 17.84 15.21	27.74 17.88 15.24	27.86 18.05 15.42	28.04 18.30 15.67	28.41 18.75 16.12	29.13 19.56 16.87

Table. Continued

fh, GHz · mm		0.1	1	3	5	7	10	15
$W/h = 6$	G	0.22 0.34 0.39	2.16 3.37 3.95	6.51 10.24 12.06	10.93 17.30 20.42	15.39 24.49 28.95	22.18 35.44 41.92	33.64 53.94 63.78
	Z	24.06 15.47 13.19	24.05 15.45 13.17	24.09 15.50 13.21	24.21 15.67 13.38	24.38 15.89 13.61	24.71 16.27 13.99	25.33 16.92 14.61
$W/h = 7$	G	0.22 0.34 0.40	2.17 3.40 3.99	6.56 10.33 12.17	11.00 17.45 20.60	15.50 24.69 29.17	22.32 35.69 42.19	33.84 54.23 64.09
	Z	21.27 13.65 11.64	21.27 13.64 11.62	21.31 13.70 11.67	21.42 13.85 11.83	21.58 14.05 12.04	21.88 14.36 12.37	22.42 14.90 12.91
$W/h = 8$	G	0.22 0.34 0.40	2.19 3.42 4.02	6.60 10.41 12.27	11.06 17.57 20.74	15.59 24.84 29.35	22.44 35.88 42.40	33.99 54.45 64.32
	Z	19.08 12.23 10.42	19.08 12.23 10.41	19.12 12.28 10.46	19.23 12.42 10.61	19.38 12.59 10.80	19.65 12.86 11.09	20.13 13.31 11.57
$W/h = 10$	G	0.22 0.34 0.40	2.20 3.46 4.07	6.65 10.53 12.41	11.16 17.75 20.96	15.72 25.07 29.62	22.62 36.16 42.71	34.21 54.76 64.65
	Z	15.84 10.12 8.63	15.83 10.12 8.61	15.88 10.18 8.68	15.98 10.30 8.81	16.11 10.43 8.97	16.33 10.63 9.21	16.72 10.98 9.60

A linear function can be used to approximate the numerical values of the propagation constant and wave impedance within two adjacent frequencies. This allows sufficient accuracy in calculating the elements of the scattering matrix.

3. CALCULATION OF SCATTERING MATRIX ELEMENTS FOR MULTI-CASCADE LPF

According to [12], the most convenient approach to calculating the scattering matrix of multi-cascade microwave filters is the transition to the transmission matrices calculated separately for each irregularity in the filter, followed by their multiplication and the reverse transition from the final transmission matrix to the final scattering matrix. In more detail, we consider the calculation of the scattering and transmission matrix for the ith-step transition shown in Fig. 3.

It is assumed that the source resistance $Z_{\mathrm{src}}(f) = Z_1(f)$ and the load resistance $Z_{\mathrm{LD}}(f) = Z_2(f)$ when the source is connected to the left arm and the load to the right arm. In this case, the coefficient of reflection from the first arm, S_{11}, and the coefficient of transmission from the first arm to the second arm, S_{21}, are functions of the frequency which are determined as follows:

$$S_{11}(f) = \frac{Z_2(f) - Z_1(f)}{Z_2(f) + Z_1(f)} \mathrm{e}^{-i\mathrm{G}_1(f)2l_1}, \tag{12}$$

$$S_{21}(f) = \sqrt{1 - \left|S_{11}(f)\right|^2}\, \mathrm{e}^{-i\left(\mathrm{G}_1(f)l_1 + \mathrm{G}_2(f)l_2\right)} = \frac{2\sqrt{Z_1(f)Z_2(f)}}{Z_2(f) + Z_1(f)} \mathrm{e}^{-i\left(\mathrm{G}_1(f)l_1 + \mathrm{G}_2(f)l_2\right)}. \tag{13}$$

Similar expressions can be obtained for the reflection coefficient from the second arm, S_{22}, and the transmission coefficient from the second arm to the first arm, S_{12}, by substituting 1 for 2 and 2 for 1 in expressions (12) and (13):

$$S_{22}(f) = \frac{Z_1(f) - Z_2(f)}{Z_1(f) + Z_2(f)} \mathrm{e}^{-i\mathrm{G}_2(f)2l_2}, \tag{14}$$

$$S_{12}(f)=\sqrt{1-\left|S_{22}(f)\right|^{2}}\,\mathrm{e}^{-i\left(\mathrm{G}_2(f)l_2+\mathrm{G}_1(f)l_1\right)}=$$
$$=\frac{2\sqrt{Z_2(f)Z_1(f)}}{Z_1(f)+Z_2(f)}\mathrm{e}^{-i\left(\mathrm{G}_2(f)l_2+\mathrm{G}_1(f)l_1\right)}. \qquad (15)$$

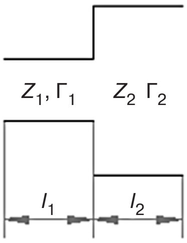


Fig. 3. The MTL step transition topology. Z_1, Z_2 are the wave impedance of the MTL segment; G$_1$, G$_2$ are the propagation constants of this segment

Thus, the scattering matrix $\mathbf{S}_i$ for the step transition can be represented by:

$$\mathbf{S}_i=\begin{bmatrix}\dfrac{Z_2(f)-Z_1(f)}{Z_2(f)+Z_1(f)}\mathrm{e}^{-i\mathrm{G}_1(f)2l_1} & \dfrac{2\sqrt{Z_2(f)Z_1(f)}}{Z_1(f)+Z_2(f)}\mathrm{e}^{-i\left(\mathrm{G}_2(f)l_2+\mathrm{G}_1(f)l_1\right)}\\ \dfrac{2\sqrt{Z_1(f)Z_2(f)}}{Z_2(f)+Z_1(f)}\mathrm{e}^{-i\left(\mathrm{G}_1(f)l_1+\mathrm{G}_2(f)l_2\right)} & \dfrac{Z_1(f)-Z_2(f)}{Z_1(f)+Z_2(f)}\mathrm{e}^{-i\mathrm{G}_2(f)2l_2}\end{bmatrix}. \qquad (16)$$

The transition from the scattering matrix $\mathbf{S}_i$ to the transmission matrix $\mathbf{T}_i$ is performed according to [12] by the following rule:

$$\mathbf{T}_i=\begin{bmatrix}T_{11}(f) & T_{12}(f)\\ T_{21}(f) & T_{22}(f)\end{bmatrix}=\begin{bmatrix}\dfrac{1}{S_{21}(f)} & -\dfrac{S_{22}(f)}{S_{21}(f)}\\ \dfrac{S_{11}(f)}{S_{21}(f)} & S_{12}(f)-\dfrac{S_{11}(f)S_{22}(f)}{S_{21}(f)}\end{bmatrix}. \qquad (17)$$

After calculating the transmission matrix $\mathbf{T}_i$ for each ith irregularity in the filter topology, the final LPF transmission matrix **T** can be obtained by multiplying the transmission matrices of all its irregularities:

$$\mathbf{T}=\prod_{i=1}^{N+1}\mathbf{T}_i. \qquad (18)$$

The transition from the transmission matrix **T** to the LPF scattering matrix **S** is performed according to [12] by the following rule:

$$\mathbf{S}=\begin{bmatrix}\dfrac{T_{21}(f)}{T_{11}(f)} & T_{22}(f)-\dfrac{T_{21}(f)T_{12}(f)}{T_{11}(f)}\\ \dfrac{1}{T_{11}(f)} & -\dfrac{T_{12}(f)}{T_{11}(f)}\end{bmatrix}. \qquad (19)$$

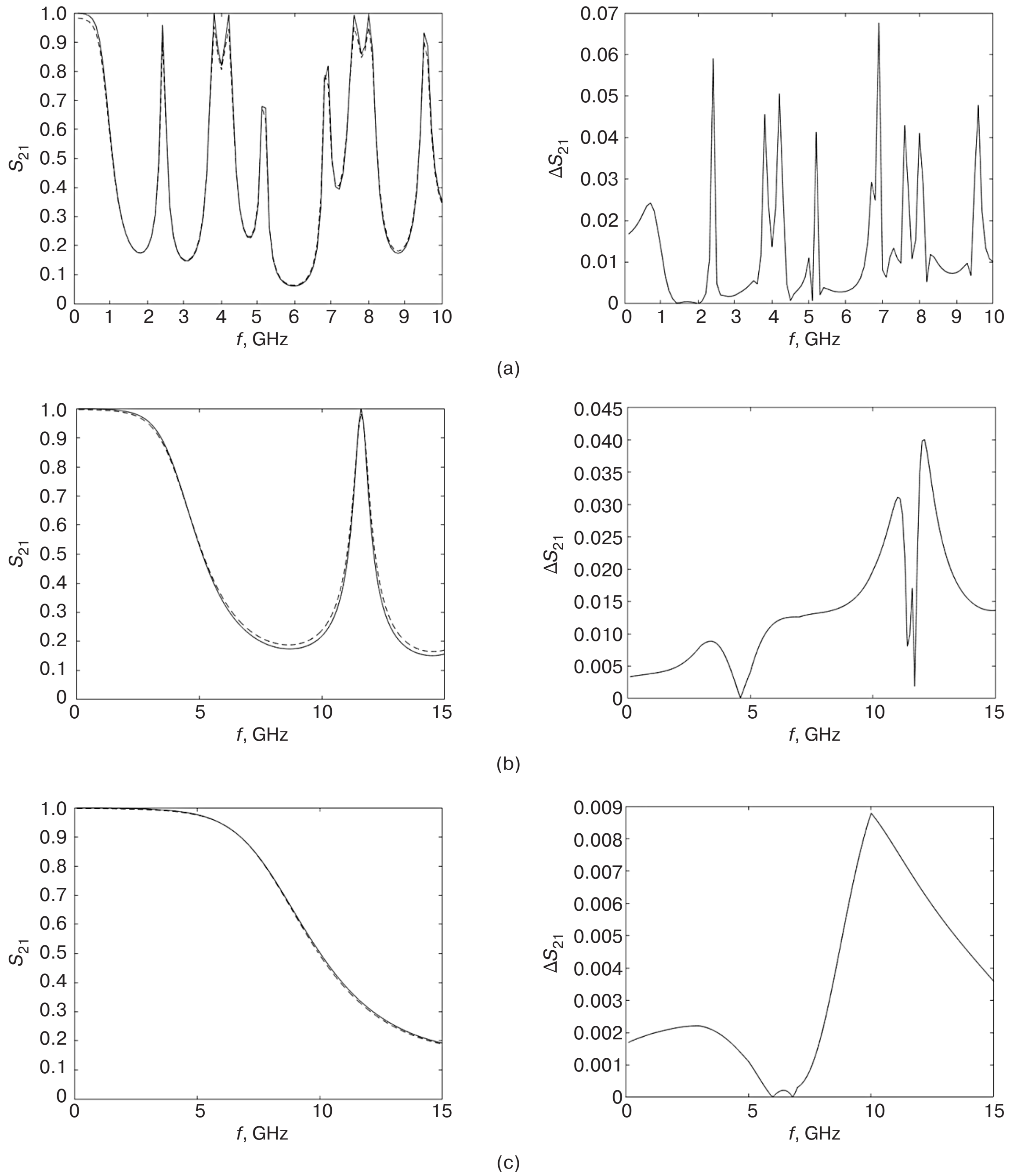
4. RESULTS OF NUMERICAL ANALYSIS

Based on the above algorithm, a software is developed using the *GNU Octave*[1] programming language to calculate the LPF characteristics in a wide range of varying strip conductor width ($0.1 \le W/h \le 10$), substrate permittivity ($2 \le \varepsilon \le 20$) and frequency (0.1 GHz $\le f \le$ 15 GHz). The structure of the software includes the main body, in which mathematical calculations are performed according to the algorithm given in this paper, as well as a subroutine for approximating the data from the table by frequency and dielectric permittivity. Based on the results obtained using the software, the LPF is calculated at different cut-off frequencies (1, 5, and 10 GHz) and its characteristics are compared with those obtained using commercial software. The amplitude-frequency characteristics of the corresponding filters (solid line is the filter

[1] https://octave.org/. Accessed March 20, 2025.

calculation using the proposed approach; dashed line is the filter calculation using commercial software) and the absolute gain calculation error values (right) are shown in Fig. 4. The time taken to calculate the transmittance using the proposed approach is only a few seconds, while calculating the transmittance using commercial software takes several minutes.

By analyzing the obtained graphs, it can be concluded that the absolute value of the transmission coefficient error does not exceed 0.08 in the whole investigated range. Simultaneously, for the LPF with 1 GHz cut-off frequency, the maximum value of the absolute transmission coefficient error is observed in the barrier band at a frequency of 7 GHz and is 0.072. For the LPF having a cut-off frequency of 5 GHz, while the maximum value of the absolute error of the transmission coefficient at a frequency of 12 GHz is 0.04. For the LPF with a cut-off frequency of 10 GHz, the maximum value of the absolute error of the transmission coefficient is 0.04.



(a)

(b)

(c)

Fig. 4. LPF amplitude frequency response at the cut-off frequency: (a) 1 GHz; (b) 5 GHz; (c) 10 GHz

CONCLUSIONS

The present work presents a microstrip LPF designed based on the MTL projection model along with a calculation of its scattering matrix. The accuracy of the obtained results is verified by comparing the transmission coefficient of the developed filter with a model constructed using modern computer-aided design systems. The absolute error of the transmission coefficient in a wide frequency band up to 15 GHz calculated based on the comparison does not exceed 0.08 for different filter cut-off frequencies. The use of the projection approach allows a significant (tenfold) reduction of the calculation time of the reflection and transmission coefficients for each pair of microwave multipole arms, together with a sufficiently high accuracy of the obtained results.

Authors' contribution. All authors equally contributed to the research work.

REFERENCES

1. Petrov I.A. Microwave filters with use broadband matching structures. *Fizika volnovykh protsessov i radiotekhnicheskie sistemy = Physics of Wave Processes and Radio Systems.* 2011;14(1):51–56 (in Russ.).
2. Lamanov Yu.A., Kudryavtseva T.O., Drobotun N.B. Design and Research Process of Microstrip Low-Pass Filters with High Slope Steepness. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki* (*Doklady TUSUR*) = *Proceedings of TUSUR University.* 2021;24(2):7–13 (in Russ.).
3. Fel'dshtein A.L. (Ed.). *Spravochnik po elementam poloskovoi tekhniki* (*Handbook of Elements of Strip Technology*). Moscow: Svjaz'; 1979. 336 p. (in Russ.).
4. Gupta K.C., Garg R., Chadha R. *Mashinnoe proektirovanie SVCh ustroistv* (*Computer-aided Design of Microwave Circuits*): transl. from Engl. Moscow: Radio i svjaz'; 1987. 432 p. (in Russ.).
   [Gupta K.C., Garg R., Chadha R. *Computer-aided Design of Microwave Circuits*. Dedham, Mass.: Artech House Inc.; 1981. 636 p.]
5. Lopatin V.V., Khvorenko V.V. Modeling and implementation of a microstrip filter. In: *Information Technologies in Science, Industry and Education: collection of papers of the Scientific and Technical Conference*. Izhevsk; 2021. P. 166–176 (in Russ.).
6. Kovalenko A.N. Projection method for constructing full-wave models of striplines. *J. Commun. Technol. Electron*. 2019;64(2):93–99. https://doi.org/10.1134/S1064226919020128
   [Original Russian Text: Kovalenko A.N. Projection method for constructing full-wave models of striplines. *Radiotekhnika i elektronika*. 2019;64(2):108–115 (in Russ.). https://doi.org/10.1134/S0033849419020128]
7. Kovalenko A.N., Yarlykov A.D. Numerical analysis of a shielded microstrip line. In: *Actual Problems and Prospects for the Development of Radio Engineering and Infocommunication Systems* (*Radioinfocom-2021*): *Proceedings of the 5th International Scientific and Practical Conference.* Moscow: RTU MIREA; 2021. P. 331–334 (in Russ.).
8. Kovalenko A.N., Yarlykov A.D. Increasing the efficiency of projection models of strip lines. *J. Commun. Technol. Electron*. 2021;66(9):997–1003. https://doi.org/10.31857/S0033849421090084
   [Original Russian Text: Kovalenko A.N., Yarlykov A.D. Increasing the efficiency of projection models of strip lines. *Radiotekhnika i elektronika*. 2021;66(9):837–844 (in Russ.). https://doi.org/10.31857/S0033849421090084]
9. Kovalenko A.N., Yarlykov A.D. Analytical expressions for electrodynamic parameters of the shielded microstrip line. *Russian Technological Journal.* 2021;9(4):68–76 (in Russ.). https://doi.org/10.32362/2500-316X-2021-9-4-68-76
10. Fusco V. *SVCh tsepi. Analiz i avtomatizirovannoe proektirovanie* (*Microwave Circuits. Analysis and Computer-aided Design*): transl. from Engl. Moscow: Radio i svyaz'; 1990. 288 p. (in Russ.).
    [Fusco V.F. *Microwave Circuits. Analysis and Computer-aided Design.* Prentice Hall; 1986. 358 p.]
11. Matthaei G.L., Yong L., Jones E.M.T. *Fil'try SVCh, soglasuyushchie tsepi i tsepi svyazi* (*Microwave Filters, Impedance-Matching Networks, and Coupling Structures*): transl. from Engl. Moscow: Svyaz'; 1972. V. 2. 496 p. (in Russ.).
    [Matthaei G.L., Yong L., Jones E.M.T. *Microwave Filters, Impedance-Matching Networks, and Coupling Structures.* McGraw-Hill; 1964. 1096 p.]
12. Kostin M.S., Yarlykov A.D. *Ustroistva i moduli sverkhvysokikh chastot* (*Devices and Modules of Ultra-High Frequencies*). Moscow, Vologda: Infra-Inzheneriya; 2022. 400 p. (in Russ.).
13. Maloratskii L.G., Yavich L.R. *Proektirovanie i raschet SVCh-elementov na poloskovykh liniyakh* (*Design and Calculation of Microwave Elements on Strip Lines*). Moscow: Sovetskoe radio; 1972. 232 p. (in Russ.).
14. Kovalenko A.N. Natural modes of a microstrip line. *Radiophys. Quantum Electron.* 1978;21(2):128–133. https://doi.org/10.1007/BF01078702
    [Original Russian Text: Kovalenko A.N. Natural modes of a microstrip line. *Izvestiya vysshikh uchebnykh zavedenii. Radiofizika*. 1978;21(2):188–194 (in Russ.).]

СПИСОК ЛИТЕРАТУРЫ

1. Петров И.А. Фильтры СВЧ с использованием широкополосных согласующих структур. *Физика волновых процессов и радиотехнические системы*. 2011;14(1):51–56.

2. Ламанов Ю.А., Кудрявцева Т.О., Дроботун Н.Б. Разработка и исследование микрополоскового фильтра низких частот с высокой крутизной спада АЧХ. *Доклады Томского государственного университета систем управления и радиоэлектроники* (*Доклады ТУСУР*). 2021;24(2):7–13. https://doi.org/10.21293/1818-0442-2021-24-2-7-13
3. *Справочник по элементам полосковой техники;* под ред. А.Л. Фельдштейна. М.: Связь; 1979. 336 с.
4. Гупта К., Гардж Р., Чадха Р. *Машинное проектирование СВЧ устройств*: пер. с англ. М.: Радио и связь; 1987. 432 с.
5. Лопатин В.В., Хворенко В.В. Моделирование и реализация микрополоскового фильтра. В сб.: *Информационные технологии в науке, промышленности и образовании: сборник трудов научно-технической конференции*. Ижевск; 2021. С. 166–176.
6. Коваленко А.Н. Проекционный метод построения электродинамических моделей полосковых линий. *Радиотехника и электроника*. 2019;64(2):108–115. https://doi.org/10.1134/S0033849419020128
7. Коваленко А.Н., Ярлыков А.Д. Численный анализ экранированной микрополосковой линии. В сб.: *Актуальные проблемы и перспективы развития радиотехнических и инфокоммуникационных систем* (*Радиоинфоком-2021*): *Сборник научных статей V Международной научно-практической конференции.* М.: РТУ МИРЭА; 2021. С. 331–334.
8. Коваленко А.Н., Ярлыков А.Д. Повышение эффективности проекционных моделей полосковых линий. *Радиотехника и электроника*. 2021;66(9):837–844. https://doi.org/10.31857/S0033849421090084
9. Коваленко А.Н., Ярлыков А.Д. Аналитические выражения для электродинамических параметров экранированной микрополосковой линии. *Russian Technological Journal.* 2021;9(4):68–76. https://doi.org/10.32362/2500-316X-2021-9-4-68-76
10. Фуско В. *СВЧ цепи. Анализ и автоматизированное проектирование*: пер. с англ. М.: Радио и связь; 1990. 288 с.
11. Маттей Д.Л., Янг Л., Джонс Е.М.Т. *Фильтры СВЧ, согласующие цепи и цепи связи*: пер. с англ. М.: Связь; 1972. Т. 1. 496 с.
12. Костин М.С., Ярлыков А.Д. *Устройства и модули сверхвысоких частот*. М., Вологда: Инфра-Инженерия; 2022. 400 с.
13. Малорацкий Л.Г., Явич Л.Р. *Проектирование и расчет СВЧ-элементов на полосковых линиях.* М.: Советское радио; 1972. 232 с.
14. Коваленко А.Н. Собственные волны микрополосковой линии. *Изв. вузов. Радиофизика.* 1978;21(2):188–194.

About the Authors

Alexey D. Yarlykov, Senior Lecturer, Department of Radio Wave Processes and Technologies, Institute of Radio Electronics and Informatics, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: yarlykov@mirea.ru. Scopus Author ID 57290652000, RSCI SPIN-code 3450-1587, https://orcid.org/0000-0002-7232-8588

Oleg A. Demin, Assistant, Department of Radio Wave Processes and Technologies, Institute of Radio Electronics and Informatics, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: demin_o@mirea.ru. https://orcid.org/0000-0002-9864-5338

Об авторах

Ярлыков Алексей Дмитриевич, старший преподаватель, кафедра радиоволновых процессов и технологий, Институт радиоэлектроники и информатики, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: yarlykov@mirea.ru. Scopus Author ID 57290652000, SPIN-код РИНЦ 3450-1587, https://orcid.org/0000-0002-7232-8588

Демин Олег Александрович, ассистент, кафедра радиоволновых процессов и технологий, Институт радиоэлектроники и информатики, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: demin_o@mirea.ru. https://orcid.org/0000-0002-9864-5338

Translated from Russian into English by K. Nazarov
Edited for English language and spelling by Thomas A. Beavitt

Micro- and nanoelectronics. Condensed matter physics

Микро- и наноэлектроника. Физика конденсированного состояния

UDC 531.8; 681.518.3
https://doi.org/10.32362/2500-316X-2025-13-3-103-121
EDN KBLENU

REVIEW ARTICLE

Microelectromechanical systems for improved gyroscope design

Pavel S. Kuznetsov @

State Scientific Research Institute of Instrument Engineering, Moscow, 129226 Russa
@ *Corresponding author, e-mail: ps_kuznetsov@mail.ru*

• **Submitted:** 31.10.2024 • **Revised:** 13.02.2025 • **Accepted:** 21.03.2025

Abstract

Objectives. Microsystem engineering is currently receiving a great deal of research attention due to the very wide scope of application of its various elements. The present study of the development and creation of modern gyroscopes based on microelectromechanical systems (MEMS gyroscopes) analyzes the risks associated with the technological aspects of their production and identifies promising areas for further development both of MEMS gyroscopes themselves and the technologies used to manufacture them.

Methods. A detailed analysis of existing scientific publications, analytical reviews, and other available sources on MEMS gyroscopes and current trends in the field of microoptoelectromechanical technologies and ferroelectric films was carried out.

Results. A brief description of the design solutions of modern MEMS gyroscopes and their integration into mechatronic systems is presented. The production technologies of MEMS gyroscopes and specifics of the technological equipment used are considered. A separate section discusses the configuration and calibration aspects of these devices. Promising directions for the development of MEMS gyroscopes with an emphasis on the use of microoptoelectromechanical converters and ferroelectric films are highlighted.

Conclusions. Based on the analysis, the prospects for the development of MEMS gyroscopes are shown, despite the existing technological challenges. It is noted that new physical principles and unique technologies can contribute to the emergence of new types of MEMS gyroscopes using micro-optoelectromechanical converters and ferroelectric films. This, in turn, opens up new horizons for future developments in this area. The necessity of developing new production technologies and specialized equipment to improve the quality of MEMS gyroscopes is demonstrated.

Keywords: MEMS gyroscope, microsystem technology, creation technology, production equipment, microoptoelectromechanical converter, optical tunneling effect, photonics, ferroelectricity

For citation: Kuznetsov P.S. Microelectromechanical systems for improved gyroscope design. *Russian Technological Journal.* 2025;13(3):103–121. https://doi.org/10.32362/2500-316X-2025-13-3-103-121, https://www.elibrary.ru/KBLENU

Financial disclosure: The author has no financial or proprietary interest in any material or method mentioned.

The author declares no conflicts of interest.

© P.S. Kuznetsov, 2025

ОБЗОРНАЯ СТАТЬЯ

Микроэлектромеханические системы: путь к совершенствованию гироскопов

П.С. Кузнецов [@]

АО «Государственный научно-исследовательский институт приборостроения», Москва, 129226 Россия
[@] Автор для переписки, e-mail: ps_kuznetsov@mail.ru

• **Поступила:** 31.10.2024 • **Доработана:** 13.02.2025 • **Принята к опубликованию:** 21.03.2025

Резюме

Цели. Микросистемная техника является одним из наиболее популярных и перспективных направлений, которые активно развиваются в настоящее время. Область применения элементов микросистемной техники весьма широка. Настоящая работа направлена на всестороннее изучение процессов разработки и создания современных гироскопов на основе микроэлектромеханических систем (МЭМС-гироскопов). Целью исследования является анализ рисков, связанных с технологическими аспектами их производства, а также определение перспективных направлений для дальнейшего развития как самих МЭМС-гироскопов, так и технологий их изготовления.

Методы. В ходе работы осуществлен детализированный анализ существующих научных публикаций, аналитических обзоров и других доступных источников, посвященных МЭМС-гироскопам и актуальным трендам в области микрооптоэлектромеханических технологий и сегнетоэлектрических пленок.

Результаты. Представлено краткое описание конструктивных решений современных МЭМС-гироскопов, а также их интеграция в мехатронные системы. Рассматриваются технологии производства МЭМС-гироскопов и специфика используемого технологического оборудования. В отдельном разделе обсуждаются аспекты настройки и калибровки этих устройств. Выделены перспективные направления развития МЭМС-гироскопов с акцентом на применение микрооптоэлектромеханических преобразователей и сегнетоэлектрических пленок.

Выводы. На основе проведенного анализа показана перспективность развития МЭМС-гироскопов, несмотря на имеющиеся технологические вызовы. Отмечено, что новые физические принципы и уникальные технологии могут способствовать появлению новых видов МЭМС-гироскопов, использующих микрооптоэлектромеханические преобразователи и сегнетоэлектрические пленки. Это, в свою очередь, открывает новые горизонты для будущих разработок в данной области. Показана необходимость разработки новых технологий производства и специализированного оборудования для повышения качества МЭМС-гироскопов.

Ключевые слова: МЭМС-гироскоп, микросистемная техника, технология создания, оборудование производства, микрооптоэлектромеханический преобразователь, оптический туннельный эффект, фотоника, сегнетоэлектричество

Для цитирования: Кузнецов П.С. Микроэлектромеханические системы: путь к совершенствованию гироскопов. *Russian Technological Journal*. 2025;13(3):103–121. https://doi.org/10.32362/2500-316X-2025-13-3-103-121, https://www.elibrary.ru/KBLENU

Прозрачность финансовой деятельности: Автор не имеет финансовой заинтересованности в представленных материалах или методах.

Автор заявляет об отсутствии конфликта интересов.

INTRODUCTION

Microsystem technology (MST) is a popular and promising area of research due to the wide field of application of its components. These include primary information sensors of electrical and non-electrical quantities, micromotors and various elements of avionics, as well as medical microinstruments.

Microelectromechanical systems (MEMS) and MEMS gyroscopes in particular are among the most demanded areas in microsystems technology [1]. The main purpose of a MEMS gyroscope is to determine the motion parameters of systems and devices in which they are installed. Their small overall dimensions and low power consumption make them attractive for use in many industries such as automotive technology, robotics, cell phones, and many others. In particular, their use in special-purpose systems (unmanned aerial vehicles, guided projectiles, inertial navigation systems, etc.) determines the increased requirements pertaining to the characteristics and technology of MEMS gyroscopes [2–6].

The purpose of the present work is to study the process of creating MEMS gyroscopes, to analyze its features, as well as to identify promising directions for the development of MEMS gyroscopes and methods of their manufacture.

SPECIFIC FEATURES OF MODERN MEMS DESIGN

Microsystems technology comprises a set of scientific, technical and technological methods that ensure the creation of an ordered composition of micron and submicron regions of materials with a given composition, structure and geometry in the volume and (or) on the surface of a solid body. This characteristic enables the realization of the functions of perception, transformation, storage, processing, translation of information, energy, motion and generation of control actions in the required modes and operating conditions [7].

A microelectromechanical system is a system that combines microelectronic and micromechanical elements (Fig. 1). These devices with the help of mechanical elements convert the external impact into an electrical signal (gyroscopes, accelerometers, pressure sensors, etc.) or under the influence of electrical forces they themselves make movements [8].

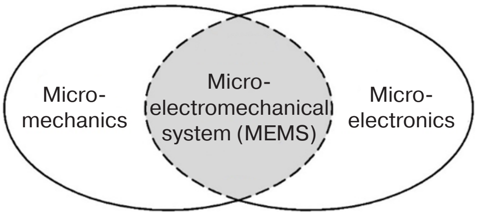


Fig. 1. MEMS diagram [8]

A MEMS gyroscope is a microminiature electromechanical system in which the energy of primary (forced) oscillations of inertial mass under the influence of external angular velocity is converted into the energy of secondary oscillations on which basis information about the measured impact may be obtained [9–11].

Thus, the simplest MEMS-based gyroscope consists of two main functional elements: an angular velocity sensor (AVS) and service electronics that perceive, amplify and processes the signal from the capacitive output of the sensor, as well as controlling the operation of the micromechanical structure. Let us consider the structural characteristics of these two elements, as well as how they are arranged in the assembly.

Most MEMS gyroscopes belong to the gyroscopes of the oscillation type. Depending on the type of inertial mass, all designs of micromechanical sensors (MMS) used in gyroscopes can be divided into several main types as presented in Fig. 2 [12, 13].

The first type is beam inertial masses. The principle of their operation can be described as follows: piezo elements provide an oscillatory motion to the cantilever beam in the direction of the *X* axis (Fig. 3). Rotation about the *Z* axis, which is parallel to the longitudinal axis of the beam, causes oscillations along the *Y* axis according to the Coriolis force, which are registered by other piezo elements [14].

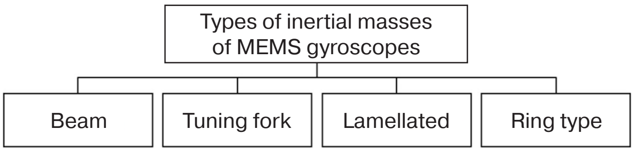


Fig. 2. Types of inertial masses of MEMS gyroscopes

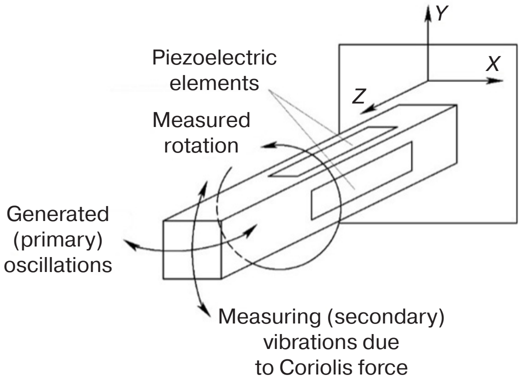


Fig. 3. Principle of operation of the beam gyroscope

The second type of inertial masses are tuning fork gyroscopes, named according to the design of the resonator. Gyroscopes built on this principle work quite simply (Fig. 4): rotation around a vertical axis causes the masses oscillating in counter-phase in one plane to oscillate in a plane perpendicular to the primary oscillations. Secondary oscillations detected using capacitive sensors provide information about the angular velocity [14].

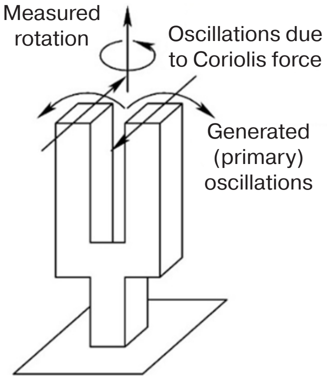

Measured rotation
Oscillations due to Coriolis force
Generated (primary) oscillations

Fig. 4. Principle of operation of the tuning fork gyroscope

Due to the presence of vertical oscillations, these gyroscopes cannot be fabricated using planar technology, which prevents their mass production.

Another type of gyroscopes utilizes plate inertial masses [9–11, 15]. Depending on the type of motion of primary and secondary oscillations of inertial masses, gyroscopes are L-L-type (linear-linear), R-R-type (rotate-rotate), and R-L-type (rotate-linear), and R-L and L-R combinations are possible (Fig. 5) [15, 16]. Significant progress in the field of L-L-type gyroscopes was achieved by Analog Devices (USA), which created the integrated MEMS technology [5]. The MEMS gyroscope of this type (Fig. 6) works as follows. Inertial masses *1*, suspended through two-dimensional springs *2*, sway in opposite directions (antiphase), causing primary oscillations to arise. The springs *2* ensure the movement of the inertial masses in two directions by means of an electrostatic force sensor. When angular velocity occurs, Coriolis forces are generated, which cause the inertial masses to move in a direction perpendicular to the direction of primary oscillation, also counter-phase. The inertial masses cause the movement of the removal combs connected with them through two-dimensional springs *2* and hung on one-dimensional springs *4*, which enable the removal brushes to move only in one direction. The take-off combs are connected with each other according to the differential scheme, which allows us to obtain at the output a signal equivalent to the acting angular velocity.

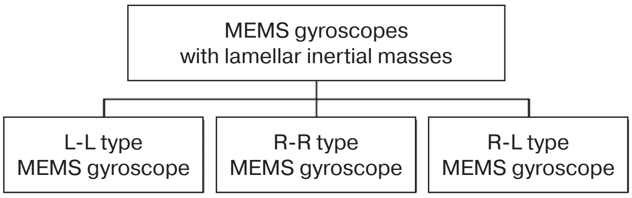


Fig. 5. Types of MEMS gyroscopes based on inertial masses

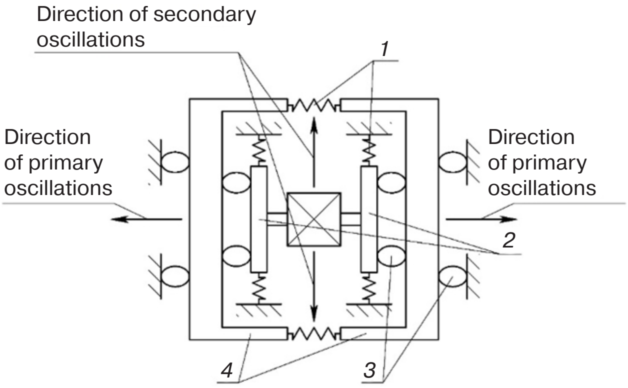

Direction of secondary oscillations
1
Direction of primary oscillations
Direction of primary oscillations
2
4
3

Fig. 6. Operating principle of L-L-type gyroscope: (*1*) one-dimensional springs; (*2*) removal strip contacts; (*3*) two-dimensional springs; (*4*) inertial masses

A schematic diagram of the R-R-type gyroscope, another type of MEMS gyroscope, is shown in Fig. 7. The inertial mass (rotor) relative to the anchors installed on the substrate (base) has a suspension mechanism including elastic and intermediate elements. The electrostatic actuator causes primary oscillatory motion of the rotor around the *Z*-axis. When the transfer angular velocity Ω_x of the base appears, the variable gyroscopic torque causes secondary oscillations of the rotor around the *Y* axis, which can be detected by capacitive displacement meters [9–11, 17, 18].

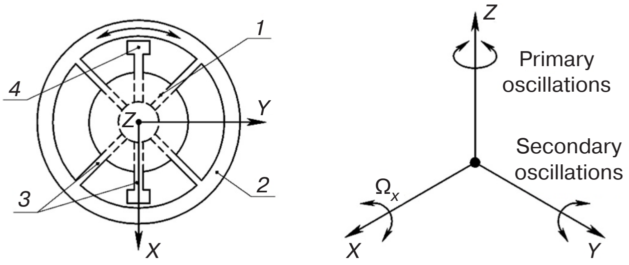


Fig. 7. Operating principle of R-R-type gyroscope: (*1*) intermediate (kinematic) suspension element; (*2*) rotor; (*3*) elastic suspension elements; (*4*) anchor

The last type of MEMS gyroscopes is the R-L-type gyroscope. According to the design (Fig. 8), it is a tuning gyroscope realized as two inertial masses fixed by elastic elements on the outer frame. The frame itself is connected to the base through elastic elements that provide it with rotational motion around the axis. The electrostatic motor, which is presented in the form of a crested structure, excites antiphase progressive oscillations of the masses. In the presence of angular velocity Ω, whose vector coincides with the measuring axis of the frame rotation, Coriolis forces arise to create a variable torque that leads to angular oscillations of the frame around the axis having a frequency equal to that of the motor. The amplitude of the frame oscillation is

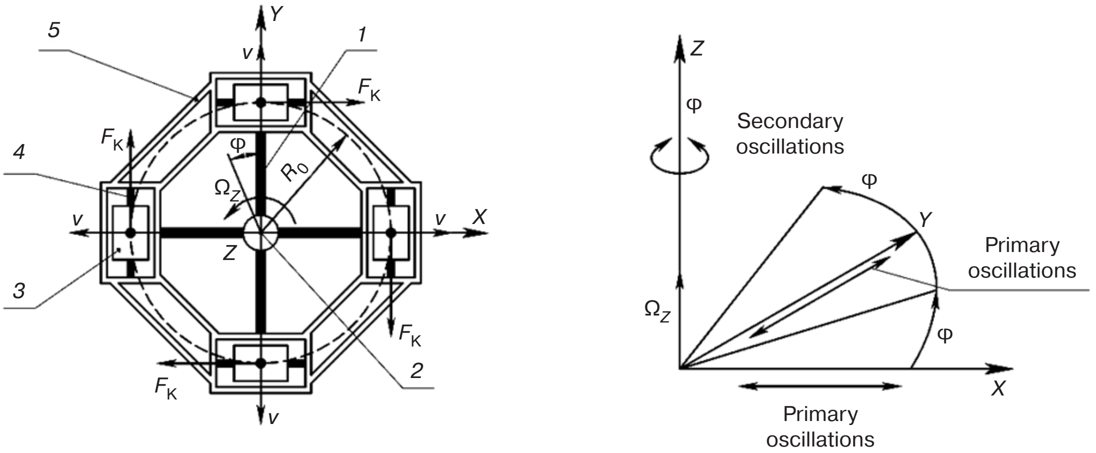


Fig. 8. Operating principle of R-L-type gyroscope:
(*1*) elastic suspension elements of secondary oscillations; (*2*) anchors; (*3*) inertial mass;
(*4*) elastic suspension elements of primary oscillations; (*5*) rigid suspension elements.
F_K is a Coriolis force vector; *v* is a velocity vector; φ is a rotation angle of the sensitive element (SE)

a measure of the angular velocity being measured. The frame oscillations are measured by means of a capacitive sensor, the electrodes of which are located on the base under the inertial masses.

The ring MEMS gyroscope is a special case of an AVS with distributed parameters. The ring resonator oscillates in the direction corresponding to the main vibrational mode. Under the influence of angular velocity (rotation of the ring), the orientation of the vibrational mode relative to the ring itself changes. This is due to the impetus to maintain its orientation under the action of the inertia force caused by Coriolis acceleration [11]. As such, the ring MEMS gyroscope can be considered as a type of wave solid-state gyroscope.

Among many existing technologies of MMS fabrication included in MEMS gyroscopes [5, 10, 11, 15, 16], let us consider in detail the silicon-on-glass technology. The micromechanical sensors manufactured by this technology comprise a vacuum-dense capsule in which the leads from the silicon structure elements are led hermetically through metallized holes in the glass to the surface where they are connected with contact pads. The micromechanical silicon structure in vacuum inside the capsule is a vibrating microgyroscope. The design of MMS in capsule version is presented in Fig. 9.

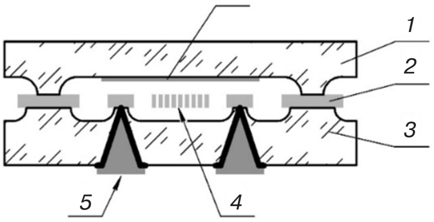


Fig. 9. MMS design in capsule version:
(*1*) cover; (*2*) base frame (Si); (*3*) base;
(*4*) silicon structure; (*5*) contacts (Al); (*6*) getter (Ti)

Several approaches are taken to the manufacture of MMS and service electronics. The smallest and the most technologically labor-intensive is the variant in which the sensor and the chip are located on one crystal and sealed in one housing. An alternative variant has similar design with the difference that the elements are executed on two different crystals. In this case, the resulting MEMS gyroscope is used as an independent element. A third design variant assumes separate encapsulation of the sensor and chip, after which they are located on the switching board together with other elements of the system. The fourth option, which is the most convenient from the point of view of its subsequent use, features a design in which the encapsulated sensor and integrated circuit are mounted on the switching board and placed in one sealed enclosure as separate elements.

Among the possible variants of MMS design, silicon-on-glass technology has the following positive features:

1) closest technology to silicon microelectronics technology and consequently well mastered;
2) technology has the possibility of group production;
3) silicon and glass wafers used in production are produced by industry;
4) specialized equipment produced for this technology is constantly upgraded and improved;
5) technology enables the production of various types of MMSs;
6) finished encapsulated element is an independent assembly element, which makes it possible to separately control its parameters, thereby reducing the yield of defective MEMS gyroscopes.

We now turn to the design of the entire MEMS gyroscope in the final design, which involves the integration of an encapsulated MMS and an integrated circuit of service electronics.

There are several variants of mutual arrangement of the capsule and chip. The first variant is a classical planar arrangement, i.e., the micromechanical converter and the service electronics circuit are located next to each other. The second variant is a two-tier arrangement, i.e., in a special case with the chip located at the bottom and the encapsulated element at the top. This arrangement has an advantage over the first embodiment because of the reduced size of the final product with a slight increase in height. However, it also requires the design of a special case, which complicates production.

The third option involves mounting the micromechanical transducer directly on the integrated circuit. The disadvantage of this design consists in the risk of damage to the chip when the capsule is mounted on it. Other design options are impossible due to the need to place the chip on the bottom for heat dissipation [19].

Let us consider in detail the second variant from the point of view of possibility of manufacturing of prototype and serial samples of MEMS gyroscopes. The finalized version (Fig. 10) with the addition of an intermediate ceramic board for mounting the micromechanical converter and its electrical switching with the microcircuit has the following advantages:

1) possibility to control all components of the MEMS gyroscope before final assembly;
2) possibility to install micromechanical transducers of various designs and sizes on the switching board;
3) sensor capsule replaceability;
4) sealing of the enclosure, providing protection from external influencing factors of the microcircuit in the enclosureless version and the micromechanical converter [20, 21].

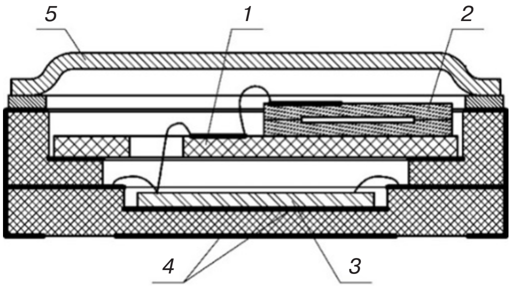


Fig. 10. MEMS gyroscope design: (*1*) switching board; (*2*) encapsulated MMS; (*3*) integrated circuit; (*4*) metallization; (*5*) cover [8]

MEMS GYROSCOPES—A CLASS OF MECHATRONIC SYSTEMS

Mechatronics is a field of science and technology based on the synergetic combination of mechanics, electronics, and a controlling computer system for the design and creation of fundamentally new systems and modules having intelligent control of their functional motion [8, 22–25]. Figure 11 depicts a schematic representation of this definition. In essence, MEMS is a mechatronic node that lacks a control system.

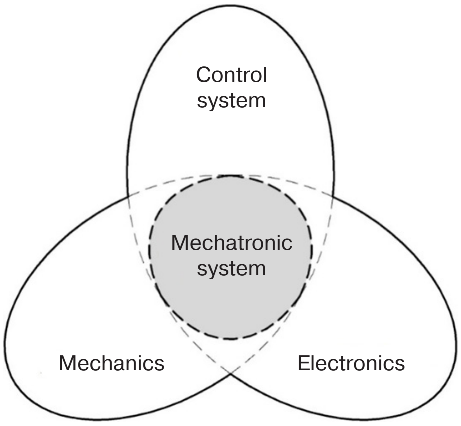


Fig. 11. Schematic diagram of the mechatronic system [8]

A more detailed study of the MEMS gyroscope design, which includes service electronics, demonstrates that it has the character of a mechatronic system. Let us analyze the products developed by GIROOPTIKA[1] (Russia) presented in Fig. 12 [14, 26–31]. Figure 13 shows the structural diagram of the micromechanical angular velocity transducer. As can be seen, in addition to the main MMD of angular velocity, the presented sensor also includes an additional MMD of linear acceleration (accelerometer), and the service electronics is represented by ASIC chip, produced by GIROOPTIKA.

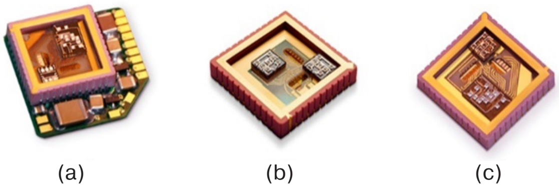


Fig. 12. Micromechanical transducers produced by GIROOPTIKA: (a) angular velocity; (b) linear acceleration; (c) complex transducer

The function of the accelerometer in the presented MEMS gyroscope is to measure linear acceleration and subsequent MMS compensation of angular velocity to accelerations.

The purpose of an ASIC chip is to provide amplification and direct digital conversion of the signal from the MMS outputs of angular velocity and linear acceleration, as well as digital formation of MEMS gyroscope output signals and control signals for MMS. In addition, the chip has a built-in processor unit with a permanent memory device (ROM), which provides the possibility

[1] http://gyro.ru/. Accessed March 22, 2025.

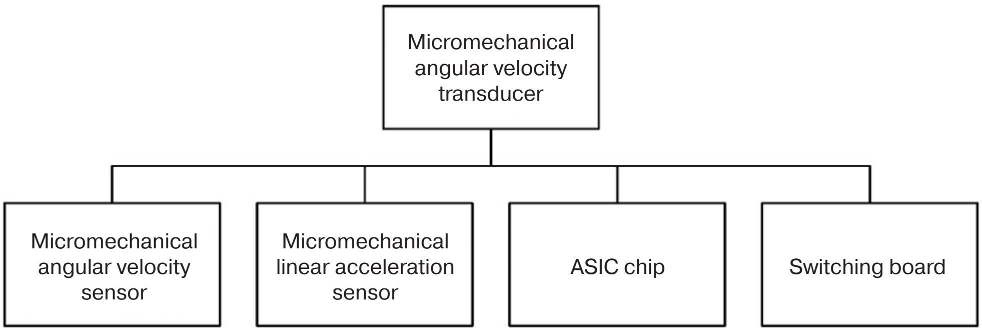

Micromechanical angular velocity transducer
Micromechanical angular velocity sensor
Micromechanical linear acceleration sensor
ASIC chip
Switching board

Fig. 13. Structural diagram of a micromechanical angular velocity transducer

of individual adjustment and calibration of each AVS, taking the technological variation of parameters and their temperature dependence into account. The processor unit is used to adjust the MMS (adjusting the frequency of natural oscillations of the MMS along the measuring axis relative to the frequency of forced oscillations), to correct the nonlinearity of the scale factor and zero offset. Compensation of technological variation of parameters and their temperature dependence and sensitivity to overload along the output axis of the angular velocity MMS is calculated in accordance with the data recorded in ROM, taking into account the signal of the on-chip built-in temperature sensor with its own analog-to-digital converter and linear acceleration MMS.

Although the complex micromechanical transducer (Fig. 12c) is similar in structure to the angular velocity transducer, the information on linear acceleration is not only used for internal correction, but is also output to an external consumer.

Thus, it is a ready-made mechatronic system capable of performing certain tasks. Further development of microsystem technology using the principles of mechatronics can lead to the creation of highly intelligent micromechatronic systems: an integrated circuit will control the entire system, micromechanical devices will, on the one hand, control and recognize the processes occurring around them, and, on the other hand, will become microminiature actuators. The first samples of micromechatronic robots already exist [8, 25, 32, 33].

BASICS OF MEMS GYROSCOPE PRODUCTION TECHNOLOGY

MEMS gyroscope manufacturing can be divided into 4 main processes:

1) MMS production in capsule version;
2) manufacturing of an integrated circuit that performs signal processing and control of the MMS;
3) switch board manufacturing;
4) finished product assembly.

The most complex processes in the creation of MEMS gyroscopes, which involve the fabrication of the MMS and the control integrated circuit, require special equipment. However, the largely typical ASIC fabrication process has already been technologically perfected. Let us dwell in more detail on the production of encapsulated MMS [34], whose manufacturing technology is based on the bulk micromechanics group technology. The deep plasma-chemical etching of silicon and anodic joining of silicon and glass wafers is necessary due to the design requiring a hermetic connection between them. In this technology, the starting materials are double-sided polished silicon wafers and glass wafers of the same diameter.

The glass plate is pre-treated, as a result of which through-holes for contacts in the lower plate are created by micro-abrasive processing, and recesses with a depth of about 50 μm are formed in the upper plate. After that, the upper plate is sprayed with a getter to maintain vacuum in the inner volume of the MMS.

The silicon wafers, which are also pre-treated, have a required silicon between 50 and 70 μm. Great care is required when handling these wafers, which are not industrially produced due to their non-standard size. Therefore, it is common to use either standard silicon wafers (100–500 μm thickness for 100 mm diameter) or silicon-on-insulator wafers with a 70 μm thick working layer and 500-μm thick silicon backing layer. This enables the use of standard equipment when processing silicon.

At the first stage of the production process cycle, photolithography and deep plasma chemical etching of silicon are performed on silicon wafers to form bilateral alignment marks on the wafer. The next step involves plasma chemical etching of cavities in the silicon. In this process, silicon oxide is used as a mask, on which photolithography is followed by etching. Next, plasma chemical etching of silicon to a depth of 20 μm is performed. Although the linear dimensions in this working layer are not critical, it is important to obtain good uniformity in depth during the etching process. After removing the silicon oxide from the formed structure, the surface is cleaned prior to carrying out the anodic bonding operation. The last process brings together the glass plate with the holes and silicon wafer. At this stage, it is important to control the gas pressure in the cavity since both low vacuum and high overpressure can lead to the destruction of the silicon layer during the subsequent thinning operation.

The essence of the silicon thinning process consist in the formation of a silicon layer having a total thickness of 70 µm from the initial silicon wafer. The silicon thinning process is followed by a projection photolithography operation. In this step, a pattern is formed with the structure in the silicon layer.

The deep plasma chemical etching operation following the projection photolithography forms the majority of the MMS structure. At an etching depth is 50 µm and minimum gap is 2 µm, the maximum aspect ratio is defined as 1 : 25. The quality of MMS functioning is affected by nonuniformity of etching and deviation of geometrical dimensions to one side or the other, as well as inclination and roughness of walls: large deviations from the set values can lead to significant deterioration of its characteristics up to rejects.

During the etching process, all the main structures of the MMS, including moving parts and elastic elements, are formed. From this point on, any operations that may damage the structure, including photoresist application and liquid plate processing, should be excluded from the technological process.

The next step in the process flow is the anodic bonding of the top glass wafer to the silicon base. At this stage, the structure is sealed at the level of the wafer. At the same time, it is necessary to maintain a given vacuum level in the MMS volume, which is achieved by thermal activation of the thin-film getter sprayed on the glass. In the process of sealing the product by anodic bonding, two opposite processes occur: the release of oxygen from the glass and its absorption by the getter at elevated temperature.

The final operations of the technological process, which are carried out at the level of the wafer, involve the creation of external metallization. At this stage, a thick layer of aluminum is sprayed, which covers the silicon contact pads at the bottom of the through holes in the glass, as well as the side walls of these holes with the metallization output to the glass surface.

The next step is the cutting of wafers into crystals, which is performed using a disk wafer cutting unit in two passes. After that, the wafer is transferred to the functional test. Those chips that successfully pass the function test are transferred to the following stages for installation into the housing and further assembly of the transducer.

When considering such technology, it is important to note that the processes of anode bonding with preliminary alignment of the wafers to be bonded (double-sided alignment) and the process of dry or deep plasma chemical etching of silicon are processes that cannot be performed on standard equipment for producing integrated circuits and/or semiconductor devices. Other processes can in principle be carried out on standard equipment with appropriate changes in the modes and materials used. This applies to chemical processing, photolithographic processes, vacuum coating processes, wafer-to-crystal separation, etc. Such a requirement may be facilitated by the use of glass and silicon wafers with standard dimensions (thickness and diameter) for the manufacture of the micromechanical elements.

Manufactured encapsulated MMSs after separation into separate crystals are checked for resonance frequencies and goodness of fit in addition to visual inspection. Here, not only the difference between the output and input frequencies will be checked, but also the presence of the necessary vacuum inside the encapsulated element. Such control permit a considerable reduction in the labor intensity and increased percentage of yield of good products at the operations of assembly and tuning of MEMS gyroscopes. However, it is impossible to completely exclude poorly working angular velocity MMSs at the early stages of manufacturing, since for this purpose it would first be necessary to connect the MMS to the processing electronics and its tuning, including mechanical effects in the form of rotations and turns.

The MEMS gyroscope is assembled by 3D-integration of an encapsulated integrated circuit and encapsulated MMS using a ceramic switching board into a special ceramic-metal housing. Integration is performed by sequential mounting of the elements into the housing followed by their mutual connection with microwires using the ball-and-wedge method.

The ASIC crystal, switch board, and encapsulated element are assembled using a conductive adhesive used in microelectronics. Sealing of the case is carried out by soldering the ceramic cover to the base of the case. A tightness check is carried out with the help of helium leak detector according to the methods and criteria used for microcircuits in ceramic-metal cases. In order to ensure that the product operates with the specified technical characteristics, transducer adjustment and calibration operations are additionally performed.

MEMS GYROSCOPE MANUFACTURING EQUIPMENT

The choice in favor of silicon-on-glass or silicon-on-insulator technologies made in the previous sections was based, among other things, on the possibility of using industrial equipment in gyroscope manufacturing technology. The selected technologies can be divided into two parts: technologies transferred from microelectronics and technologies that are inherent only in the manufacture of micromechanical devices. Thus, the equipment providing the corresponding technological processes is also divided into two groups. The equipment of the first group was initially produced by the USSR industrial sector and later by the CIS countries. However, today this market is dominated by foreign manufacturers from various countries and regions.

There is a wide range of options for selecting manual, semi-automatic or automatic equipment for standard microelectronics processes such as vacuum deposition, photolithography operations, chemical treatments, thermal oxidation, etc. Special attention should be paid to equipment designed for special processes within the selected technology of volume micromechanics and silicon-on-glass technologies. Such processes should include:

- deep plasma chemical etching of silicon and glass;
- double-sided connection of silicon and glass wafers;
- anode bonding of silicon and glass wafers without loss of alignment accuracy;
- silicon thinning on glass.

A few features of MEMS equipment should be noted here:

- specialized equipment for micromechanics operations is high-precision and very expensive and is produced by manufacturers only to order and for specific technology and customer requirements;
- the same basic equipment, as a rule, is manufactured in two modifications. The first is a variant of manual or semi-automatic equipment designed for research and development, small batch production or pilot production. The second modification is an automatic equipment with loading through cassettes, designed for manufacturing;
- the equipment is also available in cluster version to combine with units performing related operations, i.e., to create a cluster that automatically performs a whole cycle of operations;
- most manufacturers have recently started to offer the technology together with the equipment, and all manufacturers include commissioning and training in the price [35].

Regardless of the specific features of the technological process, the main requirement for equipment for the production of elements and devices of microelectronics and micromechanics consist in the possibility to maintain production with the lowest percentage of defects.

Requirements for the level of introduced contaminants and the composition of the residual gas environment inside the working chamber play an important role both in microelectronics manufacturing (neighboring tracks shorting out (Fig. 14a)) and in micromechanics manufacturing (microparticle blocking motion (Fig. 14b)). Figure 15 shows the structure of high technology equipment, where the pumping means and motion input elements in the vacuum allow us to create and maintain an ultra-clean vacuum environment. In addition, they have the ability to protect the process volume from particles and contaminants created by other elements of the vacuum system. This is primarily due to the fact that cryogenic pumping elements have no moving elements at all, while devices with contactless magnetic interaction have no rubbing elements. Their main features and basic properties are presented in numerous specialized literature [36–41].

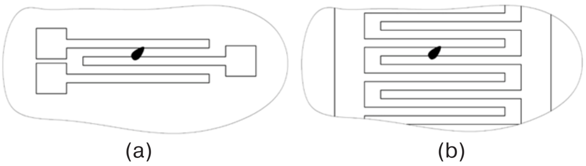
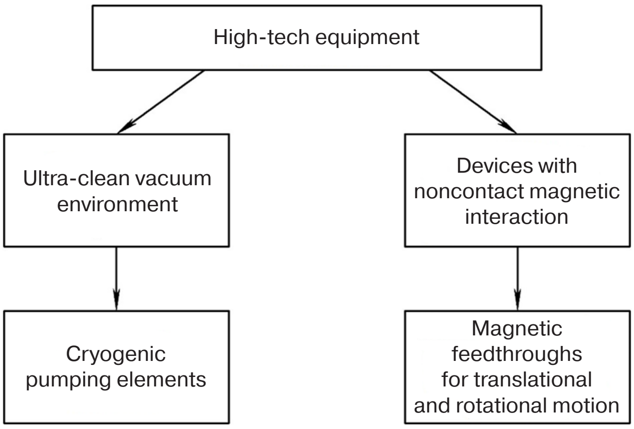


Fig. 14. Trapped microparticles on the product surface: (a) shorting of neighboring chip tracks; (b) blocking the movement of micromechanics crests


High-tech equipment

Ultra-clean vacuum environment

Devices with noncontact magnetic interaction

Cryogenic pumping elements

Magnetic feedthroughs for translational and rotational motion

Fig. 15. Structure of environmentally friendly high-tech equipment

DETERMINATION OF PARAMETERS IN THE MANUFACTURING OF MEMS GYROSCOPES

The functional purpose of MEMS angular velocity detection (MEMS gyroscopes) is to convert non-electrical physical quantities (angular velocity) into an electrical measuring signal containing quantitative information about the influencing angular velocity.

The main parameters determining the functional purpose and application area are as follows:

- angular velocity measuring range;
- resolving power;
- scaling factor nonlinearity.

The main technical characteristic of the MEMS gyroscope is the output (conversion) characteristic, i.e., the dependence of the output signal on the values of the determined angular velocities within the measurement range. The output characteristic used in the transducer channel of the MEMS gyroscope is an information channel that provides the generation of information about the angular velocity projections on the sensitivity axes of the AVS and transmission of this information to the consumer in accordance with the information exchange protocol.

Therefore, it is necessary to take into account the errors in the output characteristics of MEMS gyroscopes that may occur during their manufacture. The errors are divided into two categories: basic errors and additional errors. Basic errors are determined under normal conditions, i.e., in the absence of external influencing factors. These include nonlinearity and instability of output characteristics. The instability of the output characteristic includes the zero offset instability and the instability of the scale factor of the MEMS gyroscope.

Additional errors occur under the influence of external factors such as ambient temperature, mechanical effects, etc. Since MEMS gyroscope MMSs are a complex three-layer structure and have temperature dependence of their parameters, the temperature error of the output characteristic has the greatest influence. This is primarily due to the fact that the measuring gaps in silicon capacitors have values of 2–3 μm, while the recorded minimum displacements have values less than a nanometer. At such small values and micromechanical structural complexity, even the use of differential measurement methods cannot exclude the influence of temperature [42, 43].

In general, it is not only the micromechanical element that is temperature dependent, but also the electronics processing the signal from capacitive sensors and controlling the gyroscope operation. Therefore, it is necessary to adjust and calibrate the MEMS gyroscope.

MEMS gyroscope tuning, which is necessary for obtaining stable output parameters, consists in setting and stabilization operations within the temperature range of the bandwidth and scaling factor of the MEMS gyroscope. In addition, the temperature drift of the zero offset must be determined and compensated.

The setting operations are carried out in a climate chamber on a rotary stand. The climate chamber is used to set the temperature according to the requirements, while the stand automatically works out the specified set of angular velocities to determine the scaling factors. A temperature sensor built into the MEMS gyroscope is used to measure the temperature.

The result of tuning consists in the dependencies of coefficients responsible for bandwidth and scaling factor on the readings of the built-in temperature sensor. These dependencies, which are presented in tabular (matrix) form, are used by the control program in the piecewise linear approximation algorithm, which calculates the coefficient values for any reading of the built-in temperature sensor.

After the temperature dependence of the coefficients is added to the control program, the temperature drift of the zero offset is determined.

MEMS gyroscope tuning, which invariably precedes calibration, is intended to ensure its operability in the range of operating temperatures and angular velocities. As a result, the MEMS gyroscope can be guaranteed to have technical parameters close to the required ones. Final adjustment of parameters is carried out during calibration.

MEMS gyroscope calibration is performed to determine the output characteristics of the angular velocity transducer channels under normal conditions (basic errors) and under the influence of external factors (additional errors). Accurate calibration over the entire temperature range is typically performed during the final setup and calibration of the inertial measurement unit (IMU) into which the MEMS gyroscopes are installed. Since the IMU controller is usually much more powerful than the integrated circuit of the sensor's control electronics, it algorithmically compensates for all errors of the MEMS gyroscopes and the IMU.

ALGORITHMIC COMPENSATION OF THE MEMS GYROSCOPE ERRORS

Compensation of sensor errors for normal conditions and for each of the operating range temperatures at which the calibration is performed is performed in the IMU controller using a special control program that uses the error compensation algorithms determined during the calibration. These algorithms are based on the use of temperature dependencies of MEMS gyroscope characteristics, which are formalized in the form of tables obtained during the calibration process.

The final version of the sensor characteristics table is obtained by simulation and control of these characteristics in a special program while checking the IMU output characteristics. Modeling is performed using the files recorded during calibration and additional measurements performed following calibration. On the basis of the obtained physical values from the unit output data and their errors during modeling, a conclusion is drawn about fulfillment or non-fulfillment of the requirements to the final IMU parameters.

Compensation of nonlinearity, instability, and asymmetry of the output characteristics of the AVS is carried out using the calibration characteristics determined during calibration in the climatic chamber in the range of operating temperatures of the IMU and across the whole range of angular velocity measurement from ±0.01°/s to the maximum value according to the documentation.

The output characteristics of the AVS after calibration in the climatic chamber are presented as a piecewise linear approximation of the real dependence of the output signal on the set value of the angular velocity for the formed temperature range set in the climatic chamber.

As a result of implementation of the above algorithms, nonlinearity, instability, and asymmetry are removed from the output signal of the AVS. Thus, the scaling factor and zero offset are made identical at all temperatures and at all angular velocities in the operating range [44–48].

Now let us consider the compensation of errors of MEMS gyroscopes caused by vibration. When the MMS is mounted on a vibrating base, inertial forces caused by vibration acceleration act on the moving masses.

MMS has increased output signal noise due to the sensitivity of gyroscopes to linear overload due to finite suspension stiffness in the in-phase direction of motion of the moving masses and technological asymmetry of the suspension. In case of asymmetric sensitivity of gyroscopes to linear acceleration in case of vibration, parasitic offset of their zero signal can be formed. Nonlinearity of sensitivity to linear acceleration under the action of constant acceleration (measured or free fall acceleration) leads to the appearance of asymmetry of sensitivity to superimposed variable acceleration and, accordingly, to the constant zero offset of gyroscopes.

Additional compensation of residual errors is performed algorithmically. The influence of MMS sensitivity to linear acceleration is reduced by calibrating them taking into account the effective gravitational acceleration 1g and introducing correction factors with reference to external accelerometers.

PROMISING DIRECTIONS OF THE MEMS GYROSCOPE DEVELOPMENT

In addition to the positive qualities of MEMS gyroscopes, such as their low cost and small overall dimensions, there are also negative aspects. This type of transducer is characterized by the high instability of parameters from start to start. Zero offset of MEMS gyroscope can reach values of about 70°/h. These features, which are inherent both to Russian and foreign samples, require periodic testing and recalibration, including the possibility of self-calibration of channels during operation. All this limits the possibility of using MEMS gyroscopes in special-purpose equipment requiring high accuracy. Moreover, their use in the equipment of other classes can lead to complications and increased final costs.

The accuracy and stability of MEMS gyroscope parameters depend on how the detection of MMS micro-movements is performed. Capacitive data acquisition is most often used, i.e., the capacitance between fixed and moving parts (electrodes) of the MMS designed for this purpose changes during movement. In this case, there is a mutual influence of control and detection circuits of the useful signal of the sensor.

The accuracy of MEMS gyroscope output parameters is also significantly affected by the signal-to-noise ratio. Attempts to eliminate this problem by improving the MMS design lead to contradictions. To increase the noise immunity of the MMS, it is necessary to increase the initial capacitance. This leads to an increase in the area of the electrodes and a decrease in the gap between them, which results in increased damping of the moving parts. In order to compensate for this, it is necessary to perforate the silicon structure of the MMS, which in turn leads to a decrease in the area of the electrodes and consequent decrease in the initial capacitance.

Improved MMS parameters can be achieved with the use of optical technologies to detect micro-movements. The combined use of MEMS and microoptics can lead to synergistic effects that can solve many problems. For example, a microoptoelectromechanical (MOEM) transducer is a miniaturized device that performs measurement and subsequent processing of an optical signal when inertial mass movements occur.

Recently, the optical tunneling effect has been increasingly used in measuring devices for data acquisition. This effect is based on the process according to which light penetrates from an optically denser medium into an optically less dense medium under the condition of complete internal reflection from the interface. In this case, the electromagnetic field appearing in the optically less dense medium exponentially decays along the normal to the interface at a distance equal to the wavelength of the radiation source. Devices based on the optical tunnel effect have high resolution, low temperature error, and high noise immunity [49–53].

The principle of operation of the primary transducer made of fused quartz is based on the dependence of the light reflection coefficient of the medium-gap-medium structure on the gap size [54]. The angle of incidence of light on the boundary between the first medium and the gap (air) is chosen such that there is a complete internal reflection at a large gap value. If the gap value is comparable to the wavelength of radiation, its part passes (tunnels) through the gap into the second medium to decrease the reflection coefficient of the medium-gap-medium structure. Thus, the power of optical radiation reflected from the medium-gap-medium structure carries information about the size of the gap and, accordingly, about the nature of the object motion.

Figure 16 shows the schematic diagram of the MOEM displacement detector. The main PE of this device is a thin plate of quartz glass on which a kind of beam is cut with the help of a laser. A laser beam is directed to the end of this plate, which spreads along the plate and transfers part of its energy to photodetectors installed at some small distance from the top and bottom of the plate. Under the action of external forces, the beam bends and some difference appears between the values of energy transmitted to the photodetectors as the distance to the sensors begins to differ (Fig. 17). This difference is what is used to determine the displacement of the beam [55].

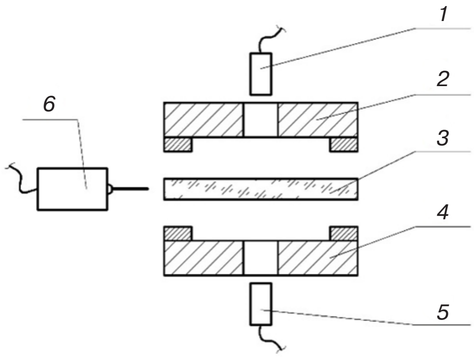


Fig. 16. Basic circuit of the MOEM detector: (*1*) photodetector F_1; (*2*) housing cover; (*3*) quartz plate (CE); (*4*) enclosure base; (*5*) photodetector F_2; (*6*) laser source [55]

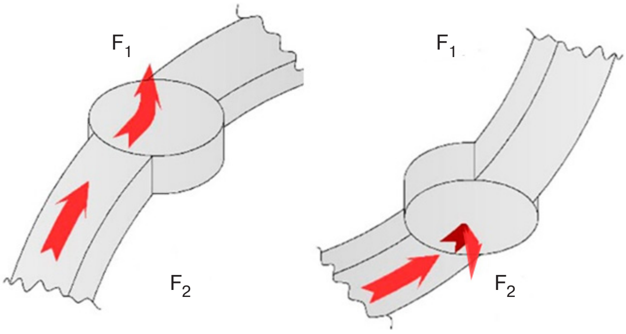


Fig. 17. Readings from the CE of the detector [55]

Technologies currently being developed involving optical processors include those aimed at optical or photonic computers, hypothetical computing devices in which calculations are performed by photons emitted by lasers or light-emitting diodes. Most current research is aimed at replacing traditional (electronic) computer components with their optical equivalents. Importantly, the frequency of a light wave is several orders of magnitude higher than the frequency of electrical signals and waves used in silicon technology. Due to the small wavelength of the light wave, it is possible to process information at increased speeds.

In most works on optical computing [56–61], the translation of information into an optical signal is required to start processing. In the design of the MOEM detector presented above (Fig. 16), there are two photodetectors for information acquisition and conversion into an electrical signal. If the laser radiation is directly sent to the optical signal receiver of the photonic calculator, the information about the CE oscillations before the processing is not transformed in any way, thus obviating the risk of distortions, which is important for special-purpose equipment.

The use of segmentoelectrics and segmentoelectric films is becoming increasingly popular in microelectronics devices, sensors, actuators, etc. Much attention is paid to the application of segmentoelectric structures in MEMS [62–66].

A segnetoelectric device is a crystalline dielectric having two or more stable (or unstable) states with different non-zero electric polarization at zero external influence (electric field, temperature, etc.), which is termed spontaneous polarization [67].

The use of segnetoelectric films in MEMS gyroscopes for motion detection has a number of advantages over classical strain-resistive and capacitive methods in terms of qualitatively expanding the capabilities of sensors. The threshold sensitivity of dynamic strain sensors based on segmentoelectric films decreases to $(\Delta l/l) \approx 10^{-9}$. The use of such devices promises to increase the sensitivity of sensors by up to two orders of magnitude as compared to existing analogs. Such generator-type sensors offer long-term stability and do not require a source of stabilized voltage.

The creation of such sensors is associated with the solution of certain technological problems. The first one consists in a combination of the technology used to create segmentoelectric films with that used for creating silicon mechanical structures. To solve the second problem, it is necessary to develop methods for creating a stable polarized state in the film [63, 68].

CONCLUSIONS

The ever-increasing demand for the production of MEMS gyroscopes and other MMSs contributes to the rapid development of microsystems technology. The entire process of creating a particular product requires constant management, not only in terms of design development, where all input elements must be precisely calculated, but also in when it comes to tuning and calibration.

An important factor in the production of such devices is the competent organization of the technological process of CE creation, which includes the operations themselves, as well as the selection and operation of special vacuum equipment.

In spite of all the discussed difficulties, novel types of MEMS gyroscopes operating on new principles are constantly appearing. This requires the development of progressive technologies for their production, as well as new specialized equipment and methods for their adjustment.

REFERENCES

1. Maltsev P.P., Telets V.A. Lecture 13. Microsystem technology. In: *Bazovye lektsii po elektronike* (*Basic Lectures on Electronics*): in 2 v. V. II. *Solid-State Electronics*. Moscow: Tekhnosfera; 2009. P. 499–575 (in Russ.).
2. Peshekhonov V.G. The Outlook for Gyroscopy. *Gyroscopy Navig*. 2020;11:193–197. https://doi.org/10.1134/S2075108720030062
3. Damianos D., Mouly J., Delbos P. *Status of the MEMS Industry 2021*. Market & Technology Report; 2021. 10 p.
4. Sinelnikov A.O., Tikhmenev N.V., Ushanov A.A., Medvedev V.M. State-of-the-Art and Development Trends of Inertial Navigation Systems Based on the Ring Laser Gyroscopes. *Fotonika = Photonics Russia*. 2024;8(6):450–466. http://doi.org/10.22184/1993-7296.FRos.2024.18.6.450.466
5. Robin L., Perlmutter M. Gyroscopes and IMUs for Defense Aerospace and Industrial. *Report by Yole Development*. 2012. 317 p.
6. Pyzhova E.M. Micromechanical gyroscopes in navigation systems. In: *Prikladnaya elektrodinamika, fotonika i zhivye sistemy – 2019* (*Applied Electrodynamics, Photonics and Living Systems – 2019*). *Proceedings.* Kazan: Individual Entrepreneur Sagieva A.R.; 2019. P. 669–671 (in Russ.). https://www.elibrary.ru/agslgx
7. Verner V.D., Ivanov A.A., Kolomenskaya N.G., Luchinin V.V., Maljtcev P.P., Popova I.V., Saurov A.N., Teletc V.A. The Produces of Microsystems Techniques – Fundamental Ideals and Terms. *Nano- i mikrosistemnaja tehnika = Nano- and Microsystem Technique.* 2007;12:2–5 (in Russ.).
8. Kuznetsov P.S. Issues and prospects for the development of mechatronics and microsystem technology. *Nano- i mikrosistemnaja tehnika = Nano- and Microsystem Technique.* 2024;26(4):159–169 (in Russ.). https://doi.org/10.17587/nmst.25.159-169
9. Raspopov V.Ya., Nikulin A.V., Likhoshurst V.V. Classification of designs of micromechanical gyros. *Izvestiya vysshikh uchebnykh zavedenii. Priborostroenie = J. Instrument. Eng.* 2005;48(8):5–8 (in Russ.).
10. Vavilov V.D., Timoshenkov S.P., Timoshenkov A.S. *Mikrosistemnye datchiki fizicheskikh velichin* (*Microsystem Sensors of Physical Quantities*). Monograph in two parts. Moscow: TEKhNOSFERA; 2018. 550 p. (in Russ.).
11. Timoshenkov S.P., Mikheev A.V., Kamenskii A.M., Artemov E.I., Polushkin V.M., Petrova N.A., Boev L.R., Puzikov V.V. Inertial microelectromechanical systems (accelerometers and gyroscopes). *Nanoindustriya = Nanoindustry*. 2020;(S96-2): 471–474 (in Russ.). https://doi.org/10.22184/1993-8578.2020.13.3s.471.474
12. Asar C., Shkel A. *MEMS Vibratory Gyroscopes. Structural Approaches to Improve Robustness.* Springer; 2009. 260 p. https://doi.org/10.1007/978-0-387-09536-3
13. Lukyanov D.P., Raspopov V.Ya., Filatov Yu.V. *Prikladnaya teoriya giroskopov* (*Applied Theory of Gyroscopes*). St. Petersburg: Electropribor; 2015. P. 42–46 (in Russ.).
14. Shakhnovich I.V. MEMS-gyroscopes – unity of choice. *Elektronika: Nauka, Tekhnologiya, Biznes = Electronics: Science, Technology, Business*. 2007;1:76–85 (in Russ.).
15. Xinfeng Z., Jingjing F., Zhipeng L., Mengtong Z. MEMS Gyroscopes Development and Application Overview on Intelligent Vehicles. In: *2020 Chinese Control and Decision Conference* (*CCDC*). Hefei, China. 2020. P. 53–59. https://doi.org/10.1109/CCDC49329.2020.9164093
16. Zhanshe G., Fucheng C., Boyu L., et al. Research development of silicon MEMS gyroscopes: a review. *Microsyst. Technol*. 2015;21(10):2053–2066. https://doi.org/10.1007/s00542-015-2645-x
17. Konovalov S.F., Ponomarev Yu.A., Mayorov D.V. Magnetic compensation of the zero signal in a two-axis hybrid R-R-R type MEMS gyro. *Vestnik Moskovskogo gosudarstvennogo tekhnicheskogo universiteta im. N.E. Baumana. Seriya Priborostroenie = Herald of the Bauman Moscow State Technical University. Series Instrument Engineering*. 2013;4(93):122–131 (in Russ.).
18. Kalikanov A.V. Calculation of the sensing element of an RR-type micromechanical gyroscope. *Molodoi uchenyi = Young Scientist*. 2020;1(291):28–33 (in Russ.).
19. Bocharov L.Yu., Maltsev P.P. Status and development prospects of microelectromechanical systems abroad. *Mikrosistemnaya tekhnika = Microsystem Technology*. 1999;1:41–46 (in Russ.).
20. Popova I.V., Lestev A.M., Semenov A.A., Ivanov V.A., Rakityanskii O.I. *Micromechanical Gyroscope-Accelerometer*: RF Pat. 84542. Publ. 10.07.2009 (in Russ.).
21. Popova I.V., Lestev A.M., Semenov A.A., Ivanov V.A., Rakityanskii O.I., Burtsev V.A. Encapsulated micromechanical gyroscopes and accelerometers for navigation and control systems. *Giroskopiya i navigatsiya = Gyroscopy and Navigation*. 2008;3(62):27–36 (in Russ.).
22. Poduraev Yu.V. Actual problems of mechatronics. *Mekhatronika, avtomatizatsiya, upravlenie = Mechatronics, Automation, Control*. 2007;4:50–53 (in Russ.).
23. Filimonov N.B. The subject of modern mechatronics and its place in the system of sciences. *Mnogoyadernye protsessory, parallel'noe programmirovanie, PLIS, sistemy obrabotki signalov = Multicore Processors, Parallel Programming, FPGAs, Signal Processing Systems*. 2017;1(7):151–159 (in Russ.).
24. Egorov O.D., Poduraev Yu.V. Analysis and synthesis of integrated mechatronic modules. *Mekhatronika, avtomatizatsiya, upravlenie = Mechatronics, Automation, Control*. 2005;10:3–8 (in Russ.).
25. Bradley D. Mechatronics – More questions than answers. *Mechatronics*. 2010;20(8):827–841. https://doi.org/10.1016/j.mechatronics.2010.07.011
26. Bogolyubov V., Bakhtieva L. Astatic Gyrocompass Based on a Hybrid Micromechanical Gyroscope. In: *IEEE East-West Design and Test Symposium,* (*EWDTS*) – *Proceedings.* 2021. P. 1–5. https://doi.org/10.1109/EWDTS52692.2021.9580982

27. Evstifeev M.I. Principal stages of development of domestic micromechanical gyroscopes. *Izvestiya vysshikh uchebnykh zavedenii. Priborostroenie = J. Instrument. Eng.* 2011;54(6):75–80 (in Russ.).
28. Evstifeev M.I. Experience in the development of micromechanical gyroscopes. In: *Navigation and Motion Control: Materials of the Reports of the 10th Anniversary Conference of Young Scientists*). St. Petersburg: Electropribor; 2008. P. 9–20 (in Russ.).
29. Moiseev N.V., Nekrasov Ya.A. Analysis of control systems of mems gyroscope channel. *Izvestiya Tul'skogo gosudarstvennogo universiteta. Tekhnicheskie nauki = News of the Tula State University. Technical Sciences*. 2012;7:115–126 (in Russ.).
30. Bekmachev A., Popova N. Inertial MEMS sensors manufactured by JSC GYROOPTIKA. *Sovremennaya elektronika = Modern Electronics*. 2018;5:4–6 (in Russ.).
31. Evstafiev S.D., Rakityanskii O.I., Severov L.A., Semenov A.A. Calibration of information characteristics MMG LL type. *Izvestiya Tul'skogo gosudarstvennogo universiteta. Tekhnicheskie nauki = News of the Tula State University. Technical Sciences.* 2012;7:167–172 (in Russ.).
32. Novikov V.Yu., Prilutskii I.S. Prospects for the development of micromechanical and nanomechanical devices. *Elektronnye sredstva i sistemy upravleniya* (*Electronic Means and Control Systems*). Materials of the Reports of the 4th International Scientific and Practical Conference. October 13–16, 2010. Tomsk: V-Spektr; 2011;2:109–113 (in Russ.).
33. Teryaev E.D., Filimonov N.B. Nanomechatronics: state, problems, prospects. *Mekhatronika, avtomatizatsiya, upravlenie = Mechatronics, Automation, Control*. 2010;1:2–14 (in Russ.).
34. Kuznetsov P.S. On the issue of technology for creating solid-state micromechanical systems. *Estestvennye i tekhnicheskie nauki = Natural and Technical Sciences*. 2024;3(190):136–139 (in Russ.). https://doi.org/10.25633/ETN.2024.03.09
35. Ivashov E.N., Kuznetsov P.S., Fedotov K.D. Optimization of control of metrological support parameters in the production of micromechanical gyroscopes. *Vestnik Mashinostroeniya*. 2015;5:40–45 (in Russ.).
36. Vasin V.A., Ivashov E.N., Kuznetsov P.S., Stepanchikov S.V. Monograph 6. Microengineering of magnetic devices. In: *Mikro- i nanoinzheneriya v elektronnom mashinostroenii* (*Micro- and Nanoengineering in Electronic Engineering*): *A series of 7 monographs*. Ivanteevka: Research Institute of Extreme Technologies; 2013. 205 p. (in Russ.).
37. Vasin V.A., Ivashov E.N., Stepanchikov S.V. Increasing the uniformity of thin film deposition in vacuum. *Tekhnologiya Mashinostroeniya*. 2011;6:27–30 (in Russ.).
38. Vasin V.A., Ivashov E.N., Kuznetsov P.S., Stepanchikov S.V. Devices with contactless magnetic interaction for special technological equipment. *Tekhnologiya Mashinostroeniya*. 2011;2:47–51 (in Russ.).
39. Vasin V.A., Ivashov E.N., Kuznetsov P.S., Stepanchikov S.V. The features of application of devices with contactless magnetic interaction in modern ultrahigh vacuum control and diagnostic and technological equipment. *Kontrol'. Diagnostika = Testing. Diagnostics*. 2011;2:44–48 (in Russ.).
40. Vasin V.A., Ivashov E.N., Stepanchikov S.V. Control and navigation system of the modern ultrahigh vacuum analytical and processing complex. *Kontrol'. Diagnostika = Testing. Diagnostics*. 2012;5:71–74 (in Russ.).
41. Ivashov E.N., Luchnikov A.P., Sigov A.S., Stepanchikov S.V. *Proektirovanie elementov i ustroistv tekhnologicheskikh sistem elektronnoi tekhniki* (*Design of Elements and Devices of Technological Systems of Electronic Equipment*). Moscow: Energoatomizdat; 2008. 288 p. (in Russ.).
42. Salenko D.S. Features of MEMS gyroscope modeling. *Avtomatika i programmnaya inzheneriya = Automatics & Software Enginery*. 2015;2(12):109–115 (in Russ.).
43. Kuznetsov P.S. Determination of parameters in the production of MEMS gyroscopes. *Innovatsii na osnove informatsionnykh i kommunikatsionnykh tekhnologii = Innovations Based on Information and Communication Technologies*. 2014;1:457–460 (in Russ.). https://www.elibrary.ru/ukrxnv
44. Kuznetsov P.S. Compensation of the nonlinearity of MEMS gyroscopes using mathematical modeling. *Innovatsii na osnove informatsionnykh i kommunikatsionnykh tekhnologii = Innovations Based on Information and Communication Technologies*. 2014;1:460–462 (in Russ.). https://www.elibrary.ru/ukrxnv
45. Boikov I.V., Krivulin N.P. Identification of discrete dynamical systems with distributed parameters. *Izvestiya vysshikh uchebnykh zavedenii. Povolzhskii region. Fiziko-matematicheskie nauk = University Proceedings. Volga region. Physical and Mathematical Sciences*. 2014;2(30):34–48 (in Russ.).
46. Krylov A.A., Korniyuk D.V. Technological approaches to zero offset compensation in MEMS gyroscopes being a part of the inertial measurement unit. *Trudy MAI*. 2018;103:18 (in Russ.).
47. Krylov A.A., Kuznetsov P.S. MEMS gyroscope zero drift elimination at different temperature dynamics. *Vestnik Kontserna VKO Almaz-Antei.* 2019;2(29):34–39 (in Russ.).
48. Krylov A.A. Research of MEMS gyroscopes zero drift instability and ways of its accounting at calibration. *Izvestiya Tul'skogo gosudarstvennogo universiteta. Tekhnicheskie nauki = News of the Tula State University. Technical Sciences.* 2020;1:64–69 (in Russ.).
49. Busurin V.I., Vasetsky S.O., Kazaryan A.V. *Compensating Micro-optoelectromechanical Angular Velocity Sensor:* RF Pat. 2806242. Publ. 30.10.2023 (in Russ.).
50. Xia D., Huang L., Zhao L. A New Design of an MOEMS Gyroscope Based on a WGM Microdisk Resonator. *Sensors.* 2019;19(12):2798. https://doi.org/10.3390/s19122798
51. Barbin E., Nesterenko T., Koleda A., Shesterikov E., Kulinich I., Kokolov A. An Optical Measuring Transducer for a Micro-Opto-Electro-Mechanical Micro-g Accelerometer Based on the Optical Tunneling Effect. *Micromachines.* 2023;14(4):802. https://doi.org/10.3390/mi14040802

52. Busurin V.I., Kazaryan A.V., Shtek S.G., Zheglov M.A., Vasetskiy S.O., Ky P.L. Frame microoptoelectromechanical angular velocity transducer with optical readout units based on optical tunneling effect. *Izmeritel'naya Tekhnika*. 2022;5:50–55 (in Russ.). https://doi.org/10.32446/0368-1025it.2022-5-50-55
53. Busurin V.I., Vasetskii S.O., Shtek S.G., Zheglov M.A. *Micro-optoelectromechanical Angular Velocity Sensor*: RF Pat. 2790042. Publ. 14.02.2023 (in Russ.).
54. Barbin E., Nesterenko T., Koleda A., Shesterikov E., Kulinich I., Kokolov A., Perin A. The Design, Modeling and Experimental Investigation of a Micro-G Microoptoelectromechanical Accelerometer with an Optical Tunneling Measuring Transducer. *Sensors*. 2024;24(3):765. https://doi.org/10.3390/s24030765
55. Shtek S.G., Zheglov M.A., Belyakov V.V., Andreasyan O.G., Vasetskiy S.O., Kuznetsov P.S. Development of a Sensitive Element of a Micro-Opto-Electromechanical Accelerometer. In: *2023 30th Saint Petersburg International Conference on Integrated Navigation Systems* (*ICINS*). https://doi.org/10.23919/ICINS51816.2023.10168511
56. Schmidt S.P., Ivanov V.V. Prospects for the creation and development of a fully optical computer. *Innovatsionnaya nauka = Innovative Science*. 2015;8-2(8):83–85 (in Russ.).
57. Molodyakov S.A. Optoelectronic processors with CCD photodetectors. Pipeline signal processing. *Informatsionno-upravlyayushchie sistemy = Information and Control systems*. 2008;6:2–8 (in Russ.).
58. Stepanenko S.A. Photonic computer: Structure and algorithms. Estimations of parameters. *Fotonika = Photonics*. 2017;7(67):72–3 (in Russ.). https://doi.org/10.22184/1993-7296.2017.67.7.72.83
59. Umarova U.B. Optical computers and their advantages. In: *Sovremennye instrumental'nye sistemy, informatsionnye tekhnologii i innovatsii* (*Modern Instrumental Systems, Information Technologies and Innovations*). Collection of Scientific Papers of the 11th International Scientific and Practical Conference: in 4 v. Kursk: Universitetskaya kniga; 2014. P. 219–220 (in Russ.).
60. Gordeev A., Voitovich V., Svyatets G. Promising photonic and phonon domestic technologies for terahertz microprocessors, RAM and interface with ultra-low power consumption. *Sovremennaya elektronika = Modern Electronics*. 2022;2:65–72 (in Russ.).
61. Rawat U., Anderson J.D., Weinstein D. Design and Applications of Integrated Transducers in Commercial CMOS Technology. *Front. Mech. Eng*. 2022;8:902421. https://doi.org/10.3389/fmech.2022.902421
62. Taishev S.R., Krainova K.Yu. Optimization of polarization conditions in nanoscale ferroelectrics for further application in sensors and MEMS. *Molodoi uchenyi = Young Scientist*. 2016;1(105):224–227 (in Russ.).
63. Vorotilov K.A., Mukhortov V.M., Sigov A.S. *Integrirovannye segnetoelektricheskie ustroistva* (*Integrated Ferroelectric Devices*). Moscow: Energoatomizdat; 2011. 175 p. (in Russ.).
64. Li S., Wang Y., Yang M., et al. Ferroelectric thin films: performance modulation and application. *Mater. Adv.* 2022;3(14): 5735–5752. https://doi.org/10.1039/D2MA00381C
65. Tadigadapa S. Piezoelectric microelectromechanical systems – challenges and opportunities. *Procedia Eng*. 2010;5:468–471. https://doi.org/10.1016/j.proeng.2010.09.148
66. Mukhortov V.M., Golovko Yu.I., Biryukov S.V., Masychev S.I., Pavlenko A.V., Stryukov D.V., Zinchenko S.P., Kovtun A.P., Tolmachev G.N. Nanoscale ferroelectric films – a new active medium for microelectronics. *Nauka Yuga Rossii = Science in the South of Russia*. 2022;18(4):33–43 (in Russ.). https://doi.org/10.7868/S25000640220404
67. Rabe K.M., Ahn C.H., Triscone J.M. (Eds.). *Fizika segnetoelektrikov: sovremennyi vzglyad* (*Physics of Ferroelectrics: A Modern Perspective*)*:* transl. from Engl. Moscow: Laboratoriya znanii; 2020. 443 p. (in Russ.). ISBN 978-5-00101-827-8 [Rabe K.M., Ahn C.H., Triscone J.M. (Eds.). *Physics of Ferroelectrics. A Modern Perspective.* Berlin, Heidelberg: Springer; 2007. 402 p.]
68. Vorotilov K.A., Sigov A.S., Romanov A.A., Mashevich P.R. Nanomaterials and structures in electronics. *Nanomaterialy i nanostruktury – 21 vek = Nanomaterials and Nanostructures – 21st Century*. 2010;1(1):45–53 (in Russ.).

СПИСОК ЛИТЕРАТУРЫ

1. Мальцев П.П., Телец В.А. Лекция 13. Микросистемная техника. В кн.: *Базовые лекции по электронике:* в 2-х т. Т. II. *Твердотельная электроника*. М.: Техносфера; 2009. С. 499–575.
2. Peshekhonov V.G. The Outlook for Gyroscopy. *Gyroscopy Navig*. 2020;11:193–197. https://doi.org/10.1134/S2075108720030062
3. Damianos D., Mouly J., Delbos P. *Status of the MEMS Industry 2021*. Market & Technology Report; 2021. 10 p.
4. Синельников А.О., Тихменев Н.В., Ушанов А.А., Медведев В.М. Современное состояние и тенденции развития инерциальных навигационных систем на кольцевых лазерных гироскопах. *Фотоника*. 2024;18(6):450–466. http://doi.org/10.22184/1993-7296.FRos.2024.18.6.450.466
5. Robin L., Perlmutter M. Gyroscopes and IMUs for Defense Aerospace and Industrial. *Report by Yole Development*. 2012. 317 p.
6. Пыжова Е.М. Микромеханические гироскопы в системах навигации. В сб.: *Прикладная электродинамика, фотоника и живые системы – 2019*: *материалы конференции*. Казань: ИП Сагиева А.Р.; 2019. С. 669–671. https://www.elibrary.ru/agslgx
7. Вернер В.Д., Сауров А.Н., Иванов А.А., Телец В.А., Коломенская Н.Г., Лучинин В.В., Мальцев П.П., Попова И.В. Изделия микросистемной техники – основные понятия и термины. *Нано- и микросистемная техника*. 2007;12:2–5.

8. Кузнецов П.С. Вопросы и перспективы развития мехатроники и микросистемной техники. *Нано- и микросистемная техника*. 2024;26(4):159–169. https://doi.org/10.17587/nmst.25.159-169
9. Распопов В.Я., Никулин А.В., Лихоширст В.В. Классификация конструкций микромеханических гироскопов. *Известия высших учебных заведений. Приборостроение*. 2005;48(8):5–8.
10. Вавилов В.Д., Тимошенков С.П., Тимошенков А.С. *Микросистемные датчики физических величин:* монография в двух частях. М.: ТЕХНОСФЕРА; 2018. 550 с.
11. Тимошенков С.П., Михеев А.В., Каменский А.М., Артемов Е.И., Полушкин В.М., Петрова Н.А., Боев Л.Р., Пузиков В.В. Инерциальные микроэлектромеханические системы (акселерометры и гироскопы). *Наноиндустрия*. 2020;(S96-2):471–474. https://doi.org/10.22184/1993-8578.2020.13.3s.471.474
12. Asar C., Shkel A. *MEMS Vibratory Gyroscopes. Structural Approaches to Improve Robustness.* Springer; 2009. 260 p. https://doi.org/10.1007/978-0-387-09536-3
13. Лукьянов Д.П., Распопов В.Я., Филатов Ю.В. *Прикладная теория гироскопов*. СПб.: ЦНИИ «Электроприбор»; 2015. С. 42–46.
14. Шахнович И.В. МЭМС-гироскопы – единство выбора. *Электроника: Наука, Технология, Бизнес*. 2007;1:76–85.
15. Xinfeng Z., Jingjing F., Zhipeng L., Mengtong Z. MEMS Gyroscopes Development and Application Overview on Intelligent Vehicles. In: *2020 Chinese Control and Decision Conference* (*CCDC*). Hefei, China. 2020. P. 53–59. https://doi.org/10.1109/CCDC49329.2020.9164093
16. Zhanshe G., Fucheng C., Boyu L., et al. Research development of silicon MEMS gyroscopes: a review. *Microsyst. Technol*. 2015;21(10):2053–2066. https://doi.org/10.1007/s00542-015-2645-x
17. Коновалов С.Ф., Пономарев Ю.А., Майоров Д.В. Магнитная компенсация нулевого сигнала в гибридном двухкоординатном МЭМС-гироскопе R-R-R-типа. *Вестник Московского государственного технического университета им. Н.Э. Баумана. Серия Приборостроение.* 2013;4(93):122–131.
18. Каликанов А.В. Расчет чувствительного элемента микромеханического гироскопа RR-типа. *Молодой ученый*. 2020;1(291):28–33.
19. Бочаров Л.Ю., Мальцев П.П. Состояние и перспективы развития микроэлектромеханических систем за рубежом. *Микросистемная техника*. 1999;1:41–46.
20. Попова И.В., Лестев А.М., Семенов А.А., Иванов В.А., Ракитянский О.И. *Микромеханический гироскоп-акселерометр*: пат. 84542 РФ. Заявка № 20009105376/22; заявл. 16.02.2009; опубл. 10.07.2009.
21. Попова И.В., Лестев А.М., Семенов А.А., Иванов В.А., Ракитянский О.И., Бурцев В.А. Капсулированные микромеханические гироскопы и акселерометры для систем навигации и управления. *Гироскопия и навигация*. 2008;3(62):27–36.
22. Подураев Ю.В. Актуальные проблемы мехатроники. *Мехатроника, автоматизация, управление*. 2007;4:50–53.
23. Филимонов Н.Б. Предмет современной мехатроники и ее место в системе наук. *Многоядерные процессоры, параллельное программирование, ПЛИС, системы обработки сигналов*. 2017;1(7):151–159.
24. Егоров О.Д., Подураев Ю.В. Анализ и синтез интегрированных мехатронных модулей. *Мехатроника, автоматизация, управление*. 2005;10:3–8.
25. Bradley D. Mechatronics – More questions than answers. *Mechatronics*. 2010;20(8):827–841. https://doi.org/10.1016/j.mechatronics.2010.07.011
26. Bogolyubov V., Bakhtieva L. Astatic Gyrocompass Based on a Hybrid Micromechanical Gyroscope. In: *IEEE East-West Design and Test Symposium,* (*EWDTS*) *– Proceedings.* 2021. P. 1–5. https://doi.org/10.1109/EWDTS52692.2021.9580982
27. Евстифеев М.И. Основные этапы разработки отечественных микромеханических гироскопов. *Известия высших учебных заведений. Приборостроение*. 2011;54(6):75–80.
28. Евстифеев М.И. Опыт разработки микромеханических гироскопов. В сб.: *Навигация и управление движением: материалы докладов X Юбилейной конференции молодых ученых*. СПб.: ЦНИИ «Электроприбор»; 2008. С. 9–20.
29. Моисеев Н.В., Некрасов Я.А. Анализ систем управления вторичными колебаниями современных микромеханических гироскопов. *Известия Тульского государственного университета. Технические науки*. 2012;7:115–126.
30. Бекмачев А., Попова Н. Инерциальные МЭМС-датчики производства АО «ГИРООПТИКА». *Современная электроника*. 2018;5:4–6.
31. Евстафьев С.Д., Ракитянский О.И., Северов Л.А., Семенов А.А. Калибровка информационных характеристик микромеханического гироскопа. *Известия Тульского государственного университета. Технические науки.* 2012;7:167–172.
32. Новиков В.Ю., Прилуцкий И.С. Перспективы развития микромеханических и наномеханических устройств. *Электронные средства и системы управления.* Материалы докладов IV Международной научно-практической конференции, 13–16 октября 2010 г. Томск: В-Спектр; 2011;2:109–113.
33. Теряев Е.Д., Филимонов Н.Б. Наномехатроника: состояние, проблемы, перспективы. *Мехатроника, автоматизация, управление*. 2010;1:2–14.
34. Кузнецов П.С. К вопросу технологии создания твердотельных микромеханических систем. *Естественные и технические науки*. 2024;3(190):136–139. https://doi.org/10.25633/ETN.2024.03.09
35. Ивашов Е.Н., Кузнецов П.С., Федотов К.Д. Оптимизация управления параметрами метрологического обеспечения при производстве микромеханических гироскопов. *Вестник машиностроения*. 2015;5:40–45.
36. Васин В.А., Ивашов Е.Н., Кузнецов П.С., Степанчиков С.В. Монография 6. Микроинженерия магнитных устройств. *Микро- и наноинженерия в электронном машиностроении: Серия из 7 монографий*. Ивантеевка Моск. обл.: Изд-во НИИ предельных технологий; 2013. 205 с.

37. Васин В.А., Ивашов Е.Н., Степанчиков С.В. Повышение равномерности нанесения тонких пленок в вакууме. *Технология машиностроения.* 2011;6:27–30.
38. Васин В.А., Ивашов Е.Н., Кузнецов П.С., Степанчиков С.В. Устройства с бесконтактным магнитным взаимодействием для специального технологического оборудования. *Технология машиностроения*. 2011;2:47–51.
39. Васин В.А., Ивашов Е.Н., Кузнецов П.С., Степанчиков С.В. Особенности применения устройств с бесконтактным магнитным взаимодействием в современном сверхвысоковакуумном контрольно-диагностическом и технологическом оборудовании. *Контроль. Диагностика*. 2011;2:44–48.
40. Васин В.А., Ивашов Е.Н., Степанчиков С.В. Контрольно-навигационная система современного сверхвысоковакуумного аналитико-технологического комплекса. *Контроль. Диагностика*. 2012;5:71–74.
41. Ивашов Е.Н., Лучников А.П., Сигов А.С., Степанчиков С.В. *Проектирование элементов и устройств технологических систем электронной техники.* М.: Энергоатомиздат; 2008. 288 с.
42. Саленко Д.С. Особенности моделирования MEMS-гироскопа. *Автоматика и программная инженерия*. 2015;2(12):109–115.
43. Кузнецов П.С. Определение параметров при производстве МЭМС-гироскопов. *Инновации на основе информационных и коммуникационных технологий.* 2014;1:456–460. https://www.elibrary.ru/ukrxnl
44. Кузнецов П.С. Компенсация нелинейности МЭМС-гироскопов с помощью математического моделирования. *Инновации на основе информационных и коммуникационных технологий.* 2014;1:460–462. https://www.elibrary.ru/ukrxnv
45. Бойков И.В., Кривулин Н.П. Идентификация дискретных динамических систем с распределенными параметрами. *Известия высших учебных заведений. Поволжский регион. Физико-математические науки*. 2014;2(30):34–48.
46. Крылов А.А., Корниюк Д.В. Технологические подходы к устранению смещения нуля МЭМС гироскопов в составе гироинерциального блока. *Труды МАИ*. 2018;103:18.
47. Крылов А.А., Кузнецов П.С. Устранение смещения нуля МЭМС-гироскопов при различной температурной динамике. *Вестник Концерна ВКО Алмаз-Антей*. 2019;2(29):34–39.
48. Крылов А.А. Исследование нестабильности дрейфа нуля МЭМС-гироскопов и способов ее учета при калибровке. *Известия Тульского государственного университета. Технические науки.* 2020;1:64–69.
49. Бусурин В.И., Васецкий С.О., Казарьян А.В. *Компенсационный микрооптоэлектромеханический датчик угловой скорости:* пат. 2806242 РФ. Заявка № 2023123201A; заявл. 06.09.2023; опубл. 30.10.2023.
50. Xia D., Huang L., Zhao L. A New Design of an MOEMS Gyroscope Based on a WGM Microdisk Resonator. *Sensors.* 2019;19(12):2798. https://doi.org/10.3390/s19122798
51. Barbin E., Nesterenko T., Koleda A., Shesterikov E., Kulinich I., Kokolov A. An Optical Measuring Transducer for a Micro-Opto-Electro-Mechanical Micro-g Accelerometer Based on the Optical Tunneling Effect. *Micromachines.* 2023;14(4):802. https://doi.org/10.3390/mi14040802
52. Бусурин В.И., Казарьян А.В., Штек С.Г., Жеглов М.А., Васецкий С.О., Чжи П.Л. Рамочный микрооптоэлектромеханический преобразователь угловой скорости с узлами оптического считывания на основе оптического туннельного эффекта. *Измерительная техника*. 2022;5:50–55. https://doi.org/10.32446/0368-1025it.2022-5-50-55
53. Бусурин В.И., Васецкий С.О., Штек С.Г., Жеглов М.А. *Микрооптоэлектромеханический датчик угловой скорости:* пат. 2790042 RF. Заявка № 2022129761; заявл. 16.11.2022; опубл. 14.02.2023.
54. Barbin E., Nesterenko T., Koleda A., Shesterikov E., Kulinich I., Kokolov A., Perin A. The Design, Modeling and Experimental Investigation of a Micro-G Microoptoelectromechanical Accelerometer with an Optical Tunneling Measuring Transducer. *Sensors*. 2024;24(3):765. https://doi.org/10.3390/s24030765
55. Shtek S.G., Zheglov M.A., Belyakov V.V., Andreasyan O.G., Vasetskiy S.O., Kuznetsov P.S. Development of a Sensitive Element of a Micro-Opto-Electromechanical Accelerometer. In: *2023 30th Saint Petersburg International Conference on Integrated Navigation Systems* (*ICINS*). https://doi.org/10.23919/ICINS51816.2023.10168511
56. Шмидт С.П., Иванов В.В. Перспективы создания и развития полностью оптического компьютера. *Инновационная наука*. 2015;8-2(8):83–85.
57. Молодяков С.А. Оптоэлектронные процессоры с ПЗС-фотоприемниками. Конвейерная обработка сигналов. *Информационно-управляющие системы*. 2008;6:2–8.
58. Степаненко С.А. Фотонный компьютер: структура и алгоритмы, оценка параметров. *Фотоника*. 2017;7(67):72–83. https://doi.org/10.22184/1993-7296.2017.67.7.72.83
59. Умарова У.Б. Оптические компьютеры и их достоинства. В сб.: *Современные инструментальные системы, информационные технологии и инновации: сборник научных трудов XI Международной научно-практической конференции:* в 4-х т. Курск: ЗАО «Университетская книга»; 2014. С. 219–220.
60. Гордеев А., Войтович В., Святец Г. Перспективные фотонные и фононные отечественные технологии для терагерцовых микропроцессоров, ОЗУ и интерфейса со сверхнизким энергопотреблением. *Современная электроника*. 2022;2:65–72.
61. Rawat U., Anderson J.D., Weinstein D. Design and Applications of Integrated Transducers in Commercial CMOS Technology. *Front. Mech. Eng*. 2022;8:902421. https://doi.org/10.3389/fmech.2022.902421
62. Таишев С.Р., Крайнова К.Ю. Оптимизация условий поляризации в наноразмерных сегнетоэлектриках с целью дальнейшего применения в датчиках и МЭМС. *Молодой ученый*. 2016;1(105):224–227.
63. Воротилов К.А., Мухортов В.М., Сигов А.С. *Интегрированные сегнетоэлектрические устройства*. М.: Энергоатомиздат; 2011. 175 с.

64. Li S., Wang Y., Yang M., et al. Ferroelectric thin films: performance modulation and application. *Mater. Adv.* 2022;3(14):5735–5752. https://doi.org/10.1039/D2MA00381C

65. Tadigadapa S. Piezoelectric microelectromechanical systems – challenges and opportunities. *Procedia Eng*. 2010;5:468–471. https://doi.org/10.1016/j.proeng.2010.09.148

66. Мухортов В.М., Головко Ю.И., Бирюков С.В., Масычев С.И., Павленко А.В., Стрюков Д.В., Зинченко С.П., Ковтун А.П., Толмачев Г.Н. Наноразмерные сегнетоэлектрические пленки – новая активная среда для микроэлектроники. *Наука Юга России*. 2022;18(4):33–43. https://doi.org/10.7868/S25000640220404

67. *Физика сегнетоэлектриков: современный взгляд;* под ред. Рабе К.М., Ана Ч.Г., Трискона Ж.М.: пер. с англ. М.: Лаборатория знаний; 2020. 443 с. ISBN 978-5-00101-827-8

68. Воротилов К.А., Сигов А.С., Романов А.А., Машевич П.Р. Формирование сегнетоэлектрических наноструктур для нового поколения устройств микро- и наноэлектроники. *Наноматериалы и наноструктуры – XXI век.* 2010;1(1):45–53.

About the Author

Pavel S. Kuznetsov, Cand. Sci. (Eng.), Deputy Head of the Experimental Complex of Microelectronics and Micromechanical Systems, State Scientific Research Institute of Instrument Engineering (GosNIIP) (125, Mira pr., Moscow, 129226 Russia). E-mail: ps_kuznetsov@mail.ru. RSCI SPIN-code 6564-9540, https://orcid.org/0000-0001-5459-7883

Об авторе

Кузнецов Павел Сергеевич, к.т.н., заместитель начальника экспериментального комплекса микроэлектроники и микромеханических систем, Акционерное общество «Государственный научно-исследовательский институт приборостроения» (АО «ГосНИИП») (129226, Россия, Москва, пр-т Мира, д. 125). E-mail: ps_kuznetsov@mail.ru. SPIN-код РИНЦ 6564-9540, https://orcid.org/0000-0001-5459-7883

Translated from Russian into English by L. Bychkova
Edited for English language and spelling by Thomas A. Beavitt

Micro- and nanoelectronics. Condensed matter physics

Микро- и наноэлектроника. Физика конденсированного состояния

UDC 537.632
https://doi.org/10.32362/2500-316X-2025-13-3-122-128
EDN IJRLMF

RESEARCH ARTICLE

Features of the magnetorefractive effect in Co–Si nanocomposites

Alexey N. Yurasov [1],
Regina Kulgunina [1],
Maxim M. Yashin [1, 2, @],
Marina A. Simdyanova [1]

[1] *MIREA – Russian Technological University, Moscow, 119454 Russia*
[2] *Bauman Moscow State Technical University, Moscow, 105005 Russia*
[@] *Corresponding author, e-mail: ihkamax@mail.ru*

• **Submitted:** 14.11.2024 • **Revised:** 03.02.2025 • **Accepted:** 20.03.2025

Abstract

Objectives. The work set out to study the spectra of the magnetorefractive effect (MRE) in the cobalt–silicon (Co–Si) nanocomposite, taking into account the contribution of the size effect (SE), and to compare the results obtained by varying the parameters of the SE. The presented approaches to investigating the magnetooptical properties of nanocomposites, which are relevant for the practical application of nondestructive testing methods, have the potential to significantly increase the efficiency of their use in various fields, including spintronics and optics.

Methods. Computer modeling approaches based on the Bruggeman approximation are used to model the examined structure as a medium with effective properties.

Results. MRE spectra obtained within the framework of the modeling fell within the range of 0.5–3.5 eV. The modeling was carried out for MRE both with and without taking into account the semiclassical size effect. The resultant modeling of the spectral dependencies of the MRE is based on the example of a Co–Si nanocomposite at different cobalt particle sizes and form factors. The influence of size effects on the form of the MRE spectra is confirmed. The reliability of the methods is confirmed by a comparison of the obtained results with empirical data. The value of the obtained results consists in the good agreement of all the calculated parameters of the discussed nanocomposite and the form of the spectral dependencies of the MRE with the results of various experiments.

Conclusions. The confirmation that both the size and form factor of granules have a significant impact on the appearance of the MRE spectra raises the prospect of developing promising nanocomposite properties at particular particle sizes. The presented results highlight the possibility of optimizing the material characteristics to improve sensitivity in magnetic sensors and noncontact devices for studying nanostructures.

Keywords: nanocomposites, effective medium theory, magnetorefractive effect, ferromagnetic, size effects

© A.N. Yurasov, R. Kulgunina, M.M. Yashin, M.A. Simdyanova, 2025

For citation: Yurasov A.N., Kulgunina R., Yashin M.M., Simdyanova M.A. Features of the magnetorefractive effect in Co–Si nanocomposites. *Russian Technological Journal.* 2025;13(3):122–128. https://doi.org/10.32362/2500-316X-2025-13-3-122-128, https://www.elibrary.ru/IJRLMF

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Особенности магниторефрактивного эффекта в нанокомпозитах Co–Si

А.Н. Юрасов [1]**,**
Р. Кулгунина [1]**,**
М.М. Яшин [1, 2, @]**,**
М.А. Симдянова [1]

[1] МИРЭА – Российский технологический университет, Москва, 119454 Россия
[2] МГТУ им. Н.Э. Баумана, Москва, 105005 Россия
[@] Автор для переписки, e-mail: ihkamax@mail.ru

• **Поступила:** 14.11.2024 • **Доработана:** 03.02.2025 • **Принята к опубликованию:** 20.03.2025

Резюме

Цели. Целью работы является исследование спектров магниторефрактивного эффекта (МРЭ) в нанокомпозитах «кобальт–кремний» (Co–Si) с учетом вклада размерного эффекта, а также сравнение полученных результатов при изменении параметров размерного эффекта. Данное исследование является важным для практического применения бесконтактных методов, т.к. оно направлено на расширение их возможностей и создание новых подходов к неразрушающему контролю и исследованию магнитооптических свойств нанокомпозитов, что может значительно повысить эффективность их использования в различных областях, включая спинтронику и оптику.

Методы. Применялось компьютерное моделирование в рамках перспективного метода эффективной среды – приближения Бруггемана, согласно которому исследуемая структура заменяется средой с эффективными свойствами.

Результаты. В рамках моделирования получены спектры МРЭ в диапазоне 0.5–3.5 эВ. При этом моделирование проводилось для МРЭ без учета и с учетом квазиклассического размерного эффекта. Конечным результатом стало моделирование спектральных зависимостей МРЭ на примере нанокомпозита Co–Si при различных значениях размера частиц и форм-фактора кобальта. Показано влияние размерных эффектов на вид спектров МРЭ. Достоверность методик хорошо подтверждается сравнением полученных результатов с эмпирическими данными, а ценность полученных результатов обусловлена тем, что все рассчитанные параметры обсуждаемого нанокомпозита и форма спектральных зависимостей МРЭ хорошо согласуются с результатами различных экспериментов.

Выводы. В рамках моделирования показано, что учет размеров и форм-фактора гранул оказывает значительное влияние на вид спектров МРЭ, демонстрируя перспективные свойства нанокомпозита при определенных размерах частиц. Представленные результаты подчеркивают возможность оптимизации характеристик материала для улучшения чувствительности в магнитных сенсорах и устройствах бесконтактного исследования наноструктур.

Ключевые слова: нанокомпозиты, теория эффективной среды, магниторефрактивный эффект, ферромагнетик, размерные эффекты

Для цитирования: Юрасов А.Н., Кулгунина Р., Яшин М.М., Симдянова М.А. Особенности магниторефрактивного эффекта в нанокомпозитах Co–Si. *Russian Technological Journal*. 2025;13(3):122–128. https://doi.org/10.32362/2500-316X-2025-13-3-122-128, https://www.elibrary.ru/IJRLMF

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

Investigating the magnetorefractive effect (MRE) in cobalt–silicon (Co–Si) nanocomposites is an important factor in the development of advanced magnetic and optical devices. Understanding the contribution of the size effect to this phenomenon is crucial for elucidating the magnetic and optical properties of nanomaterials, thus expanding the possibilities for various kinds of noncontact research [1–3].

Significant changes in the physicochemical properties of nanocomposites as the result of decreased grain size can include significant improvements in MRE functionality. In this connection, Co–Si nanocomposites are of particular interest due to their unique magneto-optical properties. A deeper understanding of the mechanisms governing the interaction between the light and the magnetic field in such systems can be achieved by modeling the MRE taking into account the size effect [4, 5].

Thus, taking into account the possible significant enhancement of practically important effects such as magnetoresistance, quantum Hall effects, MRE and many others, investigation of the properties of promising nanostructures represents an urgent task [6]. Co–Si nanocomposite materials provide an interesting example of a nanostructure; by modeling the observed optical and magneto-optical effects represents a useful noncontact and nondestructive approach to estimating characteristic parameters of the studied samples [7, 8].

MATHEMATICAL MODEL AND CALCULATION METHODS

In the paper, a mathematical model based on the effective medium theory is developed to analyze the MRE in composite materials containing cobalt nanoparticles in a silicon matrix. The main purpose of the calculation is to investigate the influence of the size effect and particle form factor on the MRE spectra.

The magnetorefractive effect describes the effect of the magnetic field on the complex refractive index of the nanocomposite, which is expressed by the change in the dielectric permittivity ε under the magnetic field [9]. In this study, the effective medium approximation (EMA) using the Bruggeman model is chosen to calculate the effective permittivity $\varepsilon^{\mathrm{EMA}}$ using the Co–Si nanocomposite as an example. The volume fraction of metallic particles (Co) in this structure is $X = 0.5$.

The calculation is carried out for particles with diameters ranging from 2 to 8 nm with different values of the form factor L in order to study the effect of varying the particle size and shape on the MRE spectra.

The magnetorefractive effect is calculated as the change in the reflection coefficient R of the nanocomposite [10]:

$$\frac{\Delta R}{R} = -(1-R)\left(\frac{\Delta\rho}{\rho}k^2\left[\frac{3n^2-k^2-1}{(n^2+k^2)((1-n)^2+k^2)}\right]\right), \qquad (1)$$

where $\frac{\Delta\rho}{\rho}$ is the magnetoresistance; k, n are the extinction and refraction coefficients, respectively.

The key parameters of the model are the diagonal and non-diagonal complex components of the dielectric permittivity tensor (DPT):

$$\gamma = \gamma_1 - i\gamma_2,\ \varepsilon = \varepsilon_{01} - i\varepsilon_{02}, \qquad (2)$$

where ε_{01} and γ_1 are the real parts of the diagonal and non-diagonal DPT components; ε_{02} and γ_2 are the imaginary parts of the DPT components, respectively.

These parameters depend on quasi-classical size effects, which are considered in the paper as a contribution of particle shape and size as captured by the MRE spectral dependence.

The size effect is accounted for by additive terms in the diagonal and non-diagonal components of the DPT based on the Drude–Lorentz model. The dielectric permittivity and the absorption coefficient of the particles are calculated with respect to the free path time τ and the concentration of the particles. The effective medium theory [11] is optimal for describing the spectral dependencies of nanostructures and nanocomposites in particular. The effective medium is described by Bruggeman's equation, which takes into account the contribution of the magnetic component of the material, the volume concentration of cobalt, and the shape of the nanoparticles:

$$X\frac{\varepsilon_1 - \varepsilon^{\mathrm{EMA}}}{\varepsilon^{\mathrm{EMA}} + L(\varepsilon_1 - \varepsilon^{\mathrm{EMA}})} + (1-X)\frac{\varepsilon_0 - \varepsilon^{\mathrm{EMA}}}{\varepsilon^{\mathrm{EMA}} + L(\varepsilon_0 - \varepsilon^{\mathrm{EMA}})} = 0, \qquad (3)$$

where ε_0 and ε_1 are the dielectric permittivities of the medium components, while L is the form factor of the medium particles.

The size effects are taken into account by varying the particle form factors L and by additives in the diagonal and non-diagonal DPT components of the nanocomposite ferromagnetic component. This is related to electron scattering on the granule surfaces. Finally, given the size effect contribution to the DPT, the DPT complex components ε_{mod} and γ_{mod} are expressed as follows, according to the Drude–Lorentz model [11, 12]:

$$\begin{aligned}\varepsilon_{\text{mod}} &= \varepsilon_{\text{Co}} + \frac{\omega_p^2}{\omega(\omega + i/\tau_{\text{bulk}})} - \frac{\omega_p^2}{\omega(\omega + i/\tau_{\text{gr}})},\\ \gamma_{\text{mod}} &= \gamma_{\text{Co}} + \frac{4\pi\sigma_{xy}^{\text{bulk}}/\tau_{\text{bulk}}^2}{\omega(\omega + i/\tau_{\text{bulk}})^2} - \frac{4\pi\sigma_{xy}^{\text{gr}}/\tau_{\text{gr}}^2}{\omega(\omega + i/\tau_{\text{gr}})^2},\end{aligned} \tag{4}$$

where ε_{Co} and γ_{Co} are the diagonal and non-diagonal components of the ferromagnetic DPT (here cobalt); ω_p is the plasma frequency; ω is the electromagnetic wave frequency; $\sigma_{xy}^{\text{bulk}} = 4\pi M_s R_{\text{bulk}} / \rho_{\text{bulk}}^2$; $\sigma_{xy}^{\text{gr}} = 4\pi M_s R_{\text{gr}} / \rho_{\text{gr}}^2$; M_s is the saturation magnetization of the ferromagnet; R_{bulk} and R_{gr} are the extraordinary Hall effect coefficients of the bulk and granules, respectively; ρ_{bulk} and ρ_{gr} are the specific resistances of the bulk and granules, respectively; τ_{bulk}, τ_{gr} are the electron mean free times in the bulk and granules, respectively.

The size effect is evident in both the extraordinary Hall effect parameter and the resistivity:

$$R_{\text{gr}} = R_{\text{bulk}} + 0.2R\frac{l}{r_0}\left(1 + \frac{l}{r_0}\right), \tag{5}$$

$$\rho_{\text{gr}} = \rho_{\text{bulk}}\left(1 + \frac{l}{r_0}\right), \tag{6}$$

where R is the value of the extraordinary Hall effect parameter of the surface material of the granules, r_0 is the particle size of the nanocomposite, and l is the electron mean free path.

MODELING RESULTS

After obtaining the values of the MRE parameter ($\Delta R/R$) ignoring the size effect by equations (1)–(3) within the framework of the promising method of the effective medium–Bruggeman approximation, the influence of the quasi-classical size effect on the spectra at different particle shape L (form factor) and cobalt particle diameter d is analyzed. A nanocomposite with cobalt volume fraction $X = 0.5$ is selected as the sample. This choice is determined by the proximity to the percolation threshold, which can significantly modify and amplify the physical effects.

As shown in Fig. 1, the most significant change in MRE observed in the nearest infrared (IR) region of the spectrum taking the size effect and particle diameter d = 2 nm into account is due to intra-band transitions. According to Fig. 2, the size effect contribution becomes noticeable only from 4 nm granule size.

In Fig. 3, the quasi-classical size effect is considered for different particle form factors. The highest effect enhancement corresponds to L = 0.2. The obtained order of magnitude results, which are in good agreement with the known experimental data for nanocomposites (e.g. [3, 4]), show a general trend of MRE enhancement with decreasing particle size and form factor.

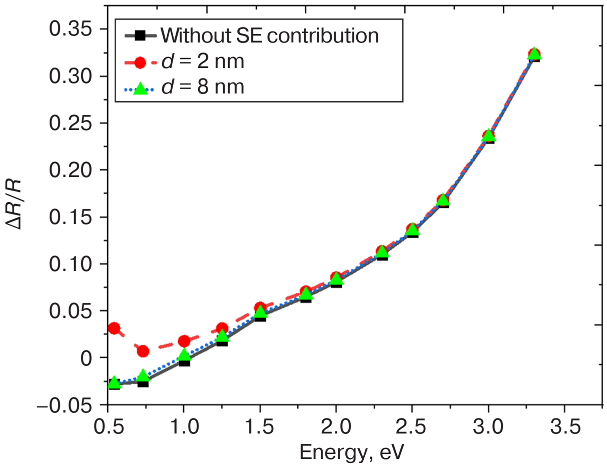


Fig. 1. MRE as a function of the incident electromagnetic wave energy without (solid line) and with the contribution of the size effect for Co particle sizes of d = 2 nm (dots) and d = 8 (dashed line) nm

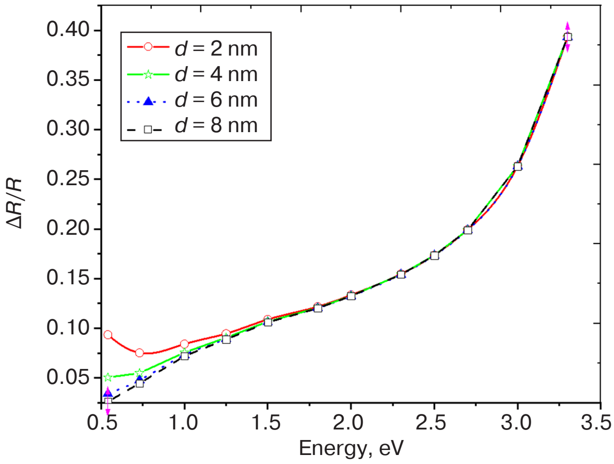


Fig. 2. MRE as a function of the incident electromagnetic wave energy considering the size effect contribution for the Co–Si nanocomposite at different particle diameters d = 2, 4, 6, and 8 nm

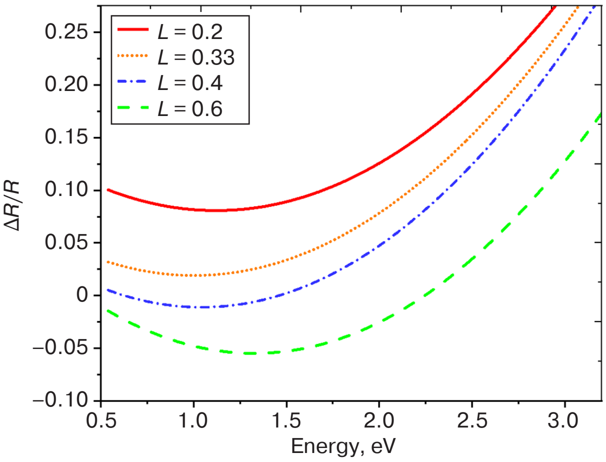


Fig. 3. MRE as a function of the incident electromagnetic wave energy considering the size effect contribution for the Co–Si nanocomposite at L = 0.2, 0.33, 0.4, and 0.6

The results open the prospect for a wide range of promising applications of nanocomposites in modern electronics based on the optical, magneto-optical, galvanomagnetic and other effects observed in the discussed nanostructures [13–15].

CONCLUSIONS

Model MRE spectra are obtained on the example of Co–Si nanocomposites taking the contribution of the size effect into account. In addition to the quasi-classical size effect, the results confirm the importance of considering the particle form factor contribution to the MRE spectral dependencies. All results are within the order of magnitude consistent with known data for similar nanostructures.

The results can be used to extend the possibilities of noncontact research methods and develop highly sensitive sensors and memory systems based on a wide range of nanostructures, such as the Co–Si nanocomposite.

ACKNOWLEDGMENTS

The study was supported by the “RTU MIREA Accelerator” program.

Authors’ contributions

A.N. Yurasov—computer simulation, discussion of results, writing and editing the text of the article.

R. Kulgunina—literature review, computer simulation, discussion of results, writing the text of the article.

M.M. Yashin—model development, computer simulation, discussion of results, writing the text of the article.

M.A. Simdyanova—literature review, computer simulation, discussion of results, writing the text of the article.

REFERENCES

1. Shkurdoda Yu.O., Dekhtyaruk L.V., Basov A.G., Chornous A.M., Shabelnyk Yu.M., Kharchenko A.P., Shabelnyk T.M. The giant magnetoresistance effect in Co/Cu/Co three-layer films. *J. Magn. Magn. Mater*. 2019;477:88–91. https://doi.org/10.1016/j.jmmm.2019.01.040
2. Jacquet J.C., Valet T. A New Magnetooptical Effect Discovered on Magnetic Multilayers: The Magnetorefractive Effect. In: *Magnetic Ultrathin Films, Multilayers and Surfaces*: *MRS Symposium Proc.* 1995. V. 384. P. 477–490. https://doi.org/10.1557/PROC-384-477
3. Granovsky A., Sukhorukov Yu., Gan’shina E., Telegin A. Magnetorefractive effect in magnetoresistive materials. In: *Magnetophotonics: From Theory to Applications.* Berlin Heidelberg: Springer; 2013. P. 107–133. http://doi.org/10.1007/978-3-642-35509-7_5
4. Kelley C.S., Naughton J., Benson E., Bradley R.C., Lazarov V.K., Thompson S.M., Matthew J.A. Investigating the magnetic field-dependent conductivity in magnetite thin films by modelling the magnetorefractive effect. *J. Phys.: Condens. Matter*. 2014;26(3):036002. http://doi.org/10.1088/0953-8984/26/3/036002
5. Krinchik G.S., Artem’ev V.A. Magneto-optical properties of Ni, Co and Fe in the ultraviolet visible and infrared parts of the spectrum. *Journal of Experimental and Theoretical Physics* (*JETF*). 1968;26(6):1080–1085.
[Original Russian Text: Krinchik G.S., Artem’ev V.A. Magneto-optical properties of Ni, Co and Fe in the ultraviolet visible and infrared parts of the spectrum. *Zhurnal Eksperimental’noi i Teorieticheskoi Fiziki*. 1967;53(6):1901–1912 (in Russ.).]
6. Yurasov A.N., Sayfulina D.A., Bakhvalova T.N. Magnetorefractive effect in metallic Co/Pt nanostructures. *Russian Technological Journal.* 2024;12(2):57–66. https://doi.org/10.32362/2500-316X-2024-12-2-57-66
7. Ganshina E., Granovsky A., Gushin V., Kuzmichov M., Podrugin P., Kravetz A., Shipil E. Optical and magneto-optical spectra of magnetic granular alloys. *Physica A: Statistical Mechanics Application.* 1997;241(1-2):45–51. https://doi.org/10.1016/S0378-4371(97)00057-5
8. Domashevskaya E.P., Ivkov S.A., Sitnikov A.V., et al. Influence of the relative content of the metal component in the dielectric matrix on the formation and size of cobalt nanocrystals in $Co_x(MgF_2)_{100-x}$ film composites. *Phys. Solid State*. 2019;61(2): 71–79. https://doi.org/10.1134/S1063783419020112

9. Reig C., Cardoso de Freitas S., Mukhopadhyay S.C. *Giant Magnetoresistance* (*GMR*) *Sensors. From Basis to State-of the-Art Applications*. *In Series: Smart Sensors, Measurement and Instrumentation*. Berlin, Heidelberg: Springer; 2013. V. 6. 301 p. ISBN 978-3-642-37172-1
10. Yurasov A.N. Magnetorefractive effect as a contactless method for investigation of functional materials. *Materialovedenie* = *Material Science*. 2014;6:32–37 (in Russ.).
11. Yurasov A.N., Yashin M.M. Methods of effective media as optimal methods for modeling the physical properties of nanostructures. *Russian Technological Journal.* 2020;8(5):68–77 (in Russ.). https://doi.org/10.32362/2500-316X-2020-8-5-68-77
12. Johnson P.B., Christy R.W. Optical constants of transition metals: Ti, V, Cr, Mn, Fe, Co, Ni, and Pd. *Phys. Rev. B.* 1974;9: 5056–5070. https://doi.org/10.1103/PhysRevB.9.5056
13. Golovan L.A., Timoshenko V.Yu., Kashkarov P.K. Optical properties of porous-system-based nanocomposites. *Phys*. *Usp*. 2007;50(6):595–612. https://doi.org/10.1070/PU2007v050n06ABEH006257

    [Original Russian Text: Golovan L.A., Timoshenko V.Yu., Kashkarov P.K. Optical properties of porous-system-based nanocomposites. *Uspekhi fizicheskikh nauk*. 2007;177(6):619–638 (in Russ.). https://doi.org/10.3367/UFNr.0177.200706b.0619]
14. Vakhrushev A.V., Fedotov A.Yu., Severyukhina O.Yu., Sidorenko A.S. Study of the influence of the cobalt structure on the magnetic properties of nanofilms. *Chemical Physics and Mesoscopy.* 2022;24(4):436–453 (in Russ.). https://doi.org/10.15350/17270529.2022.4.36
15. Demirer F.E., Lavrijsen R., Koopmans B. An investigation of the interface and bulk contributions to the magneto-optic activity in Co/Pt multi-layered thin films. *J. Appl. Phys*. 2021;129(16):163904. https://doi.org/10.1063/5.0047093

СПИСОК ЛИТЕРАТУРЫ

1. Shkurdoda Yu.O., Dekhtyaruk L.V., Basov A.G., Chornous A.M., Shabelnyk Yu.M., Kharchenko A.P., Shabelnyk T.M. The giant magnetoresistance effect in Co/Cu/Co three-layer films. *J. Magn. Magn. Mater*. 2019;477:88–91. https://doi.org/10.1016/j.jmmm.2019.01.040
2. Jacquet J.C., Valet T. A New Magnetooptical Effect Discovered on Magnetic Multilayers: The Magnetorefractive Effect. In: *Magnetic Ultrathin Films, Multilayers and Surfaces*: *MRS Symposium Proc.* 1995. V. 384. P. 477–490. https://doi.org/10.1557/PROC-384-477
3. Granovsky A., Sukhorukov Yu., Gan'shina E., Telegin A. Magnetorefractive effect in magnetoresistive materials. In: *Magnetophotonics: From Theory to Applications.* Berlin Heidelberg: Springer; 2013. P. 107–133. http://doi.org/10.1007/978-3-642-35509-7_5
4. Kelley C.S., Naughton J., Benson E., Bradley R.C., Lazarov V.K., Thompson S.M., Matthew J.A. Investigating the magnetic field-dependent conductivity in magnetite thin films by modelling the magnetorefractive effect. *J. Phys.: Condens. Matter*. 2014;26(3):036002. http://doi.org/10.1088/0953-8984/26/3/036002
5. Кринчик Г.С., Артемьев В.А. Магнитооптические свойства Ni, Co, и Fe в ультрафиолетовой, видимой и инфракрасной областях спектра. *Журнал экспериментальной и теоретической физики*. 1967;53(6):1901–1912.
6. Юрасов А.Н., Сайфулина Д.А., Бахвалова Т.Н. Магниторефрактивный эффект в металлических наноструктурах Co/Pt. *Russian Technological Journal.* 2024;12(2):57–66. https://doi.org/10.32362/2500-316X-2024-12-2-57-66
7. Ganshina E., Granovsky A., Gushin V., Kuzmichov M., Podrugin P., Kravetz A., Shipil E. Optical and magneto-optical spectra of magnetic granular alloys. *Physica A: Statistical MechanicsApplication.* 1997;241(1-2):45–51. https://doi.org/10.1016/S0378-4371(97)00057-5
8. Domashevskaya E.P., Ivkov S.A., Sitnikov A.V., et al. Influence of the relative content of the metal component in the dielectric matrix on the formation and size of cobalt nanocrystals in $Co_x(MgF_2)_{100-x}$ film composites. *Phys. Solid State*. 2019;61(2): 71–79. https://doi.org/10.1134/S1063783419020112
9. Reig C., Cardoso de Freitas S., Mukhopadhyay S.C. *Giant Magnetoresistance* (*GMR*) *Sensors. From Basis to State-of the-Art Applications*. *In Series: Smart Sensors, Measurement and Instrumentation*. Berlin, Heidelberg: Springer; 2013. V. 6. 301 p. ISBN 978-3-642-37172-1
10. Юрасов А.Н. Магниторефрактивный эффект как бесконтактный метод исследования функциональных материалов. *Материаловедение.* 2014;6:32–37.
11. Юрасов А.Н., Яшин М.М. Методы эффективной среды как оптимальные методы моделирования физических свойств наноструктур. *Russian Technological Journal.* 2020;8(5):68–77. https://doi.org/10.32362/2500-316X-2020-8-5-68-77
12. Johnson P. B., Christy R.W. Optical constants of transition metals: Ti, V, Cr, Mn, Fe, Co, Ni, and Pd. *Phys. Rev. B.* 1974;9:5056–5070. https://doi.org/10.1103/PhysRevB.9.5056
13. Головань Л.А., Тимошенко В.Ю., Кашкаров П.К. Оптические свойства нанокомпозитов на основе пористых систем. *УФН*. 2007;177(6):619–638. https://doi.org/10.3367/UFNr.0177.200706b.0619
14. Вахрушев А.В., Федотов А.Ю., Северюхина О.Ю., Сидоренко А.С. Исследование влияния структуры кобальта на магнитные свойства нанопленок. *Chemical Physics and Mesoscopy.* 2022;24(4):436–453. https://doi.org/10.15350/17270529.2022.4.36
15. Demirer F.E., Lavrijsen R., Koopmans B. An investigation of the interface and bulk contributions to the magneto-optic activity in Co/Pt multi-layered thin films. *J. Appl. Phys*. 2021;129(16):163904. https://doi.org/10.1063/5.0047093

About the Authors

Alexey N. Yurasov, Dr. Sci. (Phys.-Math.), Professor, Department of Nanoelectronics, Institute for Advanced Technologies and Industrial Programming, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: alexey_yurasov@mail.ru. ResearcherID M-3113-2016, Scopus Author ID 6602974416, RSCI SPIN-code 4259-8885, https://orcid.org/0000-0002-9104-3529

Regina Kulgunina, Student, Institute for Advanced Technologies and Industrial Programming, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: yummy2002@mail.ru. https://orcid.org/0009-0004-1864-9155

Maxim M. Yashin, Cand. Sci. (Phys.–Math.), Associate Professor, Department of Nanoelectronics, Institute for Advanced Technologies and Industrial Programming, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia); Associate Professor, Department of Physics, Bauman Moscow State Technical University (5, 2-ya Baumanskaya ul., Moscow, 105005 Russia). E-mail: ihkamax@mail.ru. ResearcherID G-6809-2017, Scopus Author ID 57210607470, RSCI SPIN-code 2438-6135, https://orcid.org/0000-0001-8022-9355

Marina A. Simdyanova, Assistant, Department of Nanoelectronics, Institute for Advanced Technologies and Industrial Programming, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: marina.simdyanova3103@mail.ru. Scopus Author ID 58532241200, RSCI SPIN-code 1447-1140, https://orcid.org/0009-0009-8418-6896

Об авторах

Юрасов Алексей Николаевич, д.ф.-м.н., профессор, кафедра наноэлектроники, Институт перспективных технологий и индустриального программирования, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: alexey_yurasov@mail.ru. ResearcherID M-3113-2016, Scopus Author ID 6602974416, SPIN-код РИНЦ 4259-8885, https://orcid.org/0000-0002-9104-3529

Кулгунина Регина, бакалавр, Институт перспективных технологий и индустриального программирования, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: yummy2002@mail.ru. https://orcid.org/0009-0004-1864-9155

Яшин Максим Михайлович, к.ф.-м.н., доцент, кафедра наноэлектроники, Институт перспективных технологий и индустриального программирования, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78); доцент, кафедра физики, ФГАОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана) (105005, Россия, Москва, ул. 2-я Бауманская, д. 5). E-mail: ihkamax@mail.ru. ResearcherID G-6809-2017, Scopus Author ID 57210607470, SPIN-код РИНЦ 2438-6135, https://orcid.org/0000-0001-8022-9355

Симдянова Марина Александровна, ассистент, кафедра наноэлектроники, Институт перспективных технологий и индустриального программирования, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: marina.simdyanova3103@mail.ru. Scopus Author ID 58532241200, SPIN-код РИНЦ 1447-1140, https://orcid.org/0009-0009-8418-6896

Translated from Russian into English by K. Nazarov
Edited for English language and spelling by Thomas A. Beavitt

MIREA – Russian Technological University.
78, Vernadskogo pr., Moscow, 119454 Russian
Federation.
Publication date May 30, 2025.
Not for sale.

МИРЭА – Российский технологический
университет.
119454, РФ, г. Москва, пр-т Вернадского, д. 78.
Дата опубликования 30.05.2025 г.
Не для продажи.

https://www.rtj-mirea.ru