2022

ANNUAL REPORT

Institute of Smart Systems and Artificial Intelligence

2022

ISSAI

CONTENTS

ISSAI

FOUNDING DIRECTOR'S MESSAGE

The beyond Turing test level interaction capability of large language models such as ChatGPT shook the world this year, making the humankind fathom the extent of activities that can be done by AI. Indeed, humanity is in the midst of its greatest disruptive economic transformation that will dwarf the industrial revolution of the 19th century. Advances in autonomous driving, smart grids, intelligent agents, 5G networks, and personalized health will have profound effects on all facets of our lives over the next decade. Artificial intelligence (AI) is a key technology for all these developments.

Countries that want to lead this race will need to reorient their economies and adopt policies to keep AI and data at the center. Training an AI workforce and establishing AI research capacity will be vital to this endeavor.

The Institute of Smart Systems and Artificial Intelligence (ISSAI) was established in 2019 on the basis of Nazarbayev University to support the AI-based development of Kazakhstan. The Institute is the first of its kind in the country and the region. It trains Kazakhstan's digital workforce by conducting advanced research projects in AI and data science. The Institute has a team of over 80 working in a gender-balanced team. The Institute's publications are already appearing in high-impact journals and at prestigious conferences. ISSAI is a strong proponent of "AI for Good" and all projects conform to the highest standards of AI ethical principles. The Institute has established an advanced computational infrastructure and shares this resource with researchers of the country.

With passing years, ISSAI has emerged as the largest repository of open-source AI datasets in Central Asia and stimulates data-based research and innovation already. ISSAI also organizes outreach programs so that a greater number of young researchers can be introduced to state-of-the-art AI techniques and hardware.

In 2022, the AI know-how of ISSAI has matured and this will serve as the foundation of groundbreaking advances in AI-based speech processing, natural language processing, biometrics, reinforcement learning, and applications of AI in construction and nutrition.

We believe that high-quality research can be conducted when interdisciplinary teams from different cultures and countries come together and unite their efforts to create a better world for future generations. Therefore, we invite everyone who shares our motto "AI for Good" to join our research activities and collaborate with us.
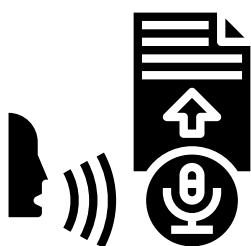
Prof. H. Atakan Varol
Founding Director of ISSAI

DATASETS

In total there are 10 ISSAI datasets in 2022

All research datasets and models developed by the Institute are open-source and publicly available for download from ISSAI website: https://issai.nu.edu.kz/issai-datasets/

KAZAKH SPEECH CORPUS 2

The first industrial-scale open-source Kazakh speech corpus (KSC2). KSC2 corpus subsumes the previously introduced two corpora: Kazakh speech corpus and Kazakh TTS 2, and supplements additional data from other sources like tv programs, radio, senate, and podcasts. In total, KSC2 contains around 1.2k hours of high-quality transcribed data comprising over 600k utterances.

KAZAKH TEXT-TO-SPEECH 2

An expanded version of the previously released Kazakh text-to-speech (KazakhTTS) synthesis corpus. In the new KazakhTTS2 corpus, the overall size has increased from 93 hours to 271 hours, the number of speakers has risen from two to five (three females and two males), and the topic coverage has been diversified with the help of new sources, including a book and Wikipedia articles.
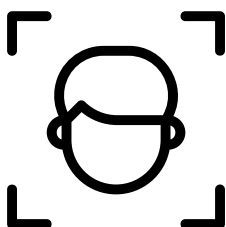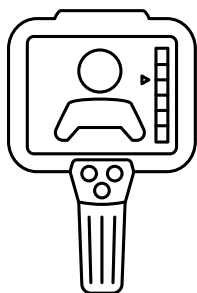
SPEAKING FACES

A large-scale publicly-available dataset designed to encourage research in the general areas of user authentication, facial recognition, speech recognition and human-computer interaction.
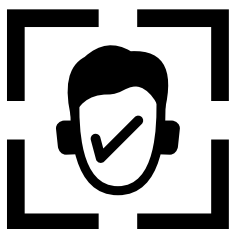
DATASETS

UZBEK SPEECH CORPUS

The USC is an open-source speech corpus that has been developed in collaboration between ISSAI and the Image and Speech Processing Laboratory in the Department of Computer Systems of the Tashkent University of Information Technologies (https://tuit.uz/en/kompyuter-tizimlari). The USC comprises 958 different speakers with a total of 105 hours of transcribed audio recordings.

TFW: ANNOTATED THERMAL FACES IN THE WILD DATASET

The dataset contains thermal images acquired in indoor (controlled) and outdoor (uncontrolled) environments. The indoor dataset was constructed using our previously published SpeakingFaces dataset. The outdoor dataset was collected using the same FLIR T540 thermal camera with a resolution of 464x348 pixels, a wave-band of 7.5–14 μm, the field of view 24, and an iron color palette. The dataset was manually annotated with face bounding boxes and five-point facial landmarks.

### SF-TL54: A THERMAL FACIAL LANDMARK DATASET WITH VISUAL PAIRS

A thermal face dataset with manually annotated bounding boxes and facial landmarks. The dataset was constructed using our large-scale SpeakingFaces dataset. In total, the dataset contains 2,556 thermal-visual image pairs of 142 subjects, where each subject has 18 thermal-visual image pairs (2 trial x 9 positions).

DATASETS

RUSSIAN SPEECH CORPUS (OPENSTT-CS334)

The OpenSTT-CS334 is a manually re-transcribed 334-hour clean subset of the Russian OpenSTT (https://github.com/snakers4/open_stt). The dataset contains recordings only from the books and YouTube domain.
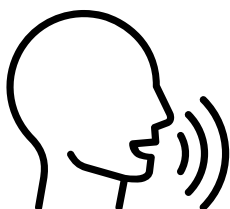
KAZAKH-ACCENTED ENGLISH

The dataset consists of Kazakh-accented English recordings (~7.7 hours) extracted from the SpeakingFaces (https://doi.org/10.48333/smgd-yj77), i.e., native Kazakh speakers uttering English verbal commands given to virtual assistants and other smart devices, such as 'turn off the lights', 'play the next song', and so on.

IMUWIFINE

A finer-level sequential dataset of IMU and WiFi received signal strengths (RSS). The dataset contains 120 trajectories covering an aggregate distance of over 14 kilometers. The dataset was collected across 3 floors of the C4 building of Nazarbayev University.

TURKIC SPEECH CORPUS

A multilingual ASR model that can recognize ten Turkic languages—Azerbaijani, Bashkir, Chuvash, Kazakh, Kyrgyz, Sakha, Tatar, Turkish, Uyghur, and Uzbek. A total of 22 models were developed (13 monolingual and 9 multilingual) using open-source datasets. The multilingual models that were trained using joint speech data performed more robustly than the baseline monolingual models, with the best model achieving an average character and word error rate reduction of 56% and 54%, respectively.

DATASETS

All research datasets and models developed by the Institute are open-source and publicly available for download from ISSAI's website. ISSAI datasets are available for free download from https://issai.nu.edu.kz/issai-datasets/

The number of downloads for each dataset

January-December 2022
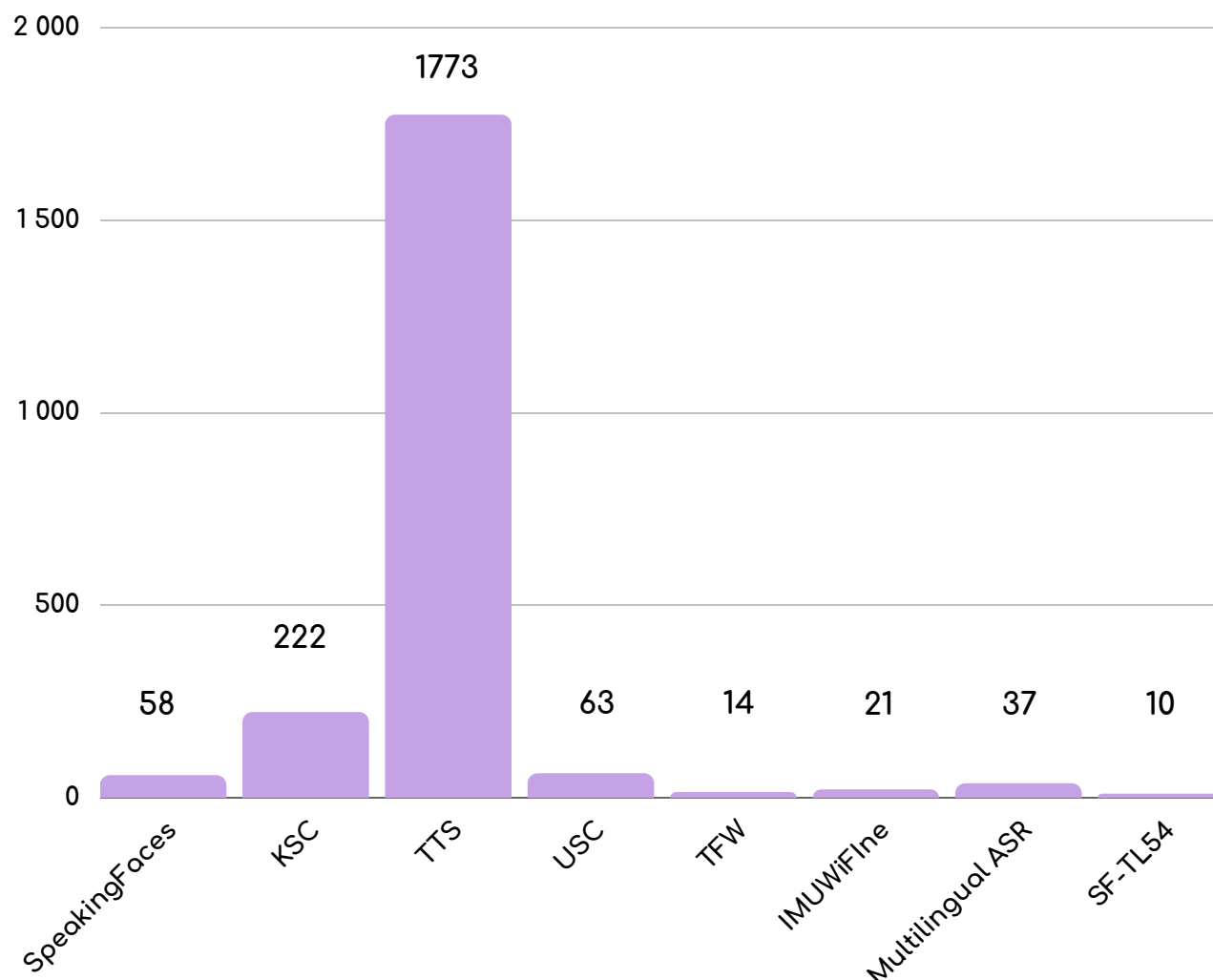
More than 100 local and international companies and educational organizations have downloaded ISSAI datasets.

The list includes affiliations like Kaspi Bank, BTS Digital, Beeline, Microsoft, Google inc, Meta, Satbayev University, Kazakh British Technical University, Astana IT University, Beijing Language and Culture University, Tianjing University, John Hopkins University, University of California Berkely, Meta etc.

ISSAI

GITHUB

GitHub open-source models:

ISSAI shares the pre-trained AI models of its projects on ISSAI GitHub page. Anyone can download the models for free.

GitHub page: https://github.com/IS2AI

32

REPOSITORIES

Top-6

REPOSITORIES

- KazakhTTS
- SpeakingFaces
- KSC
- Chest Xray module
- SF-TL54
- TFW

Top

LANGUAGES

- Python
- Shell
- Jupyter Notebook
- Java Script
- C#

ISSAI

ISSAI TEAM

ISSAI strives to be a diverse community providing opportunities for everyone and encouraging and supporting women in technical ideas. In 2022, 42 out of 82 members of ISSAI team were female.
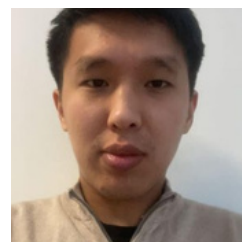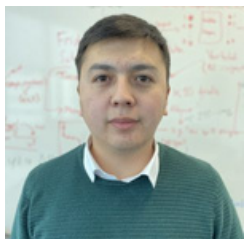
HUSEYIN ATAKAN VAROL
FOUNDING DIRECTOR

YERBOL ABSALYAMOV
TECHNICAL PROJECT COORDINATOR

KURALAY BAIMENOVA
SENIOR PR MANAGER

GIBRAT KURMANOV
ADMINISTRATIVE MANAGER

ASKAT KUZDEUOV
SENIOR DATA SCIENTIST

MADINA ABDRAKHMANOVA
DATA SCIENTIST
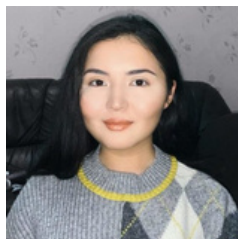
SAIDA MUSSAKHOJAYEVA
DATA SCIENTIST
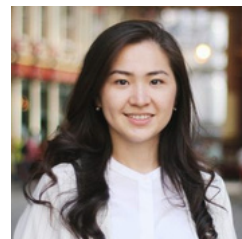
AKNUR KARABAY
DATA SCIENTIST

MAKAT TLEBALIYEV
COMPUTER ENGINEER
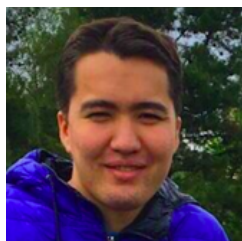
SHAKHIZAT NURGALIYEV
COMPUTER ENGINEER

RUSTEM YESHPANOV
TECHNICAL WRITER
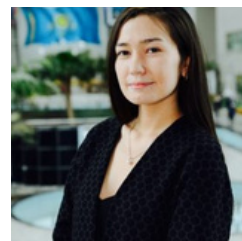
ULZHAN BISSARINOVA
DATA SCIENTIST
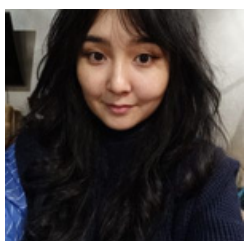
TOMIRIS RAKHIMZHANOVA
RESEARCH ASSISTANT

ZAREMA BALGABAYEVA
RESEARCH ASSISTANT

RAUSHAN UTEMURATOVA
RESEARCH ASSISTANT

TIMUR UNASPEKOV
RESEARCH ASSISTANT

ISSAI TEAM

ISSAI invites undergraduate and graduate students from diverse academic backgrounds to complete an internship. An intern works alongside data scientists and post-doctoral researchers on ISSAI projects and gains first-hand experience with state-of-the-art computational facilities. Internships can be completed on either a paid or unpaid (possibly credit-bearing) basis, depending on the applicant's educational background and skills. The average duration of an internship is six to eight weeks; the minimum duration of an internship is one month. Applicants go through a recruitment procedure including interview.

In 2022, 22% of research assistants were Bachelor students, and 19.5 % of research assistant were Master and PhD students.

Full-time workers
13.4%

other RAs
45.1%

Bachelor students
22%

Master students
12.2%

PhD students
7.3%

Female
51%

Male
49%
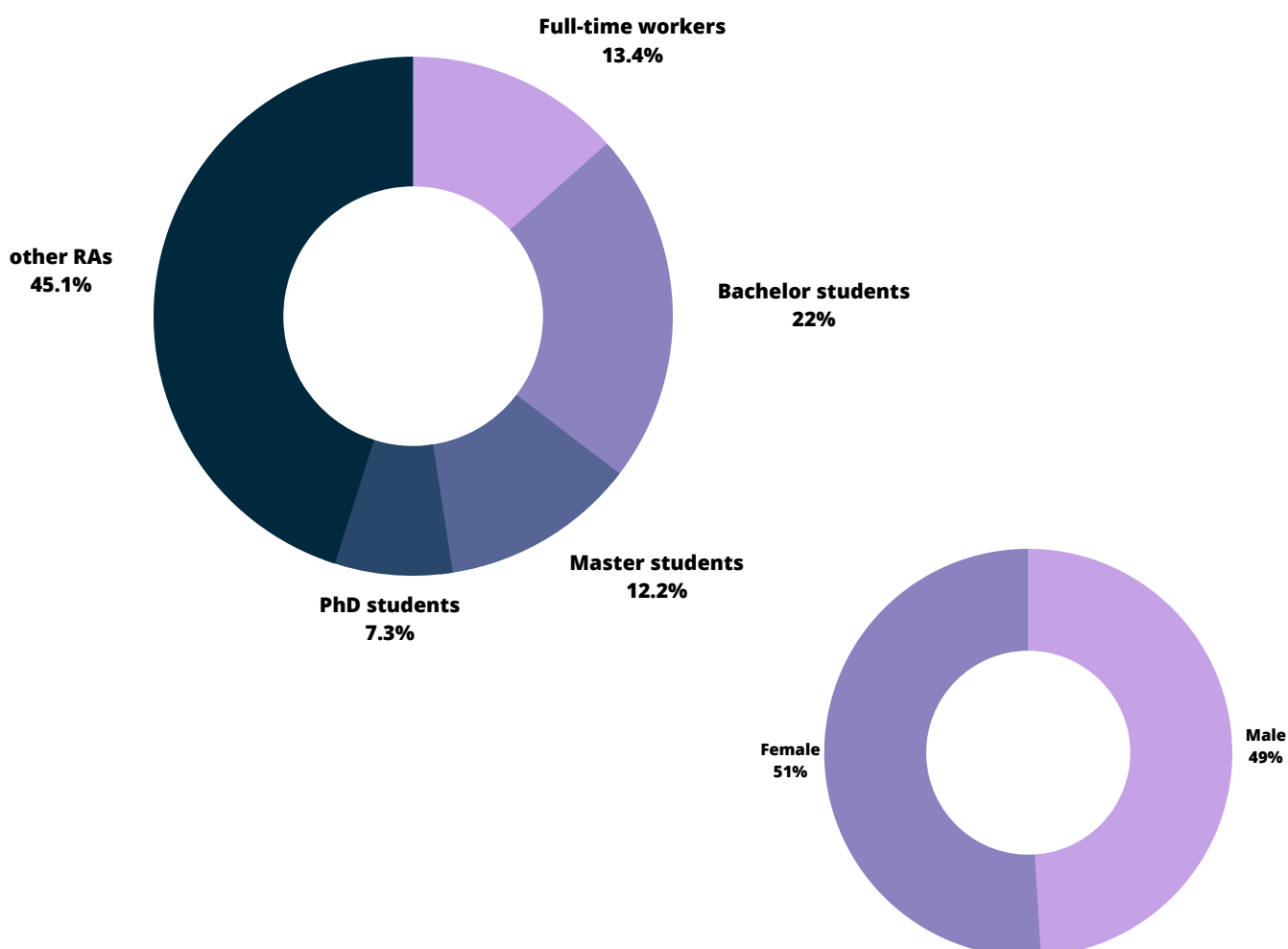
CONFERENCES

PARTICIPATION IN CONFERENCES 2022

14TH IEEE/SICE INTERNATIONAL SYMPOSIUM ON SYSTEM INTEGRATION (SII 2022)

ISSAI participated in the 14th IEEE/SICE International Symposium on System Integration (SII 2022), held online on 9-12th January 2022. At the Symposium ISSAI data scientist Askat Kuzdeuov presented two papers:

1. SF-TL54: A Thermal Facial Landmark Dataset with Visual Pairs by Askat Kuzdeuov, Darina Koishigarina, Dana Aubakirova, Saniya Abushakimova, and Huseyin Atakan Varol.
2. End-To-End Sequential Indoor Localization Using Smartphone Inertial Sensors and WiFi by Mukhamet Nurpeiissov, Askat Kuzdeuov, Aslan Assylkhanov, Yerbolat Khassanov, and Huseyin Atakan Varol.

GLOBAL YOUNG SCIENTISTS SUMMIT 2022

SSAI researcher Zhanat Makhataeva participated in the Global Young Scientists Summit 2022 (GYSS2022) from 17th to 21st January 2022. GYSS2022 brings together bright young researchers and top experts to discuss scientific and technological trends and how research can address important global challenges. The theme of the summit was "Advancing Science, Creating Technologies for a Better World". Under this theme, promising young scientists spent four days exchanging ideas and knowledge with the speakers and their peers.

ISSAI LAUNCHES THERMAL FACES IN THE WILD DATASET

A paper entitled "TFW: Annotated Thermal Faces in the Wild" by the ISSAI data scientists A. Kuzdeuov, D. Aubakirova, D. Koishigarina and H. A. Varol was accepted to the prestigious journal IEEE Transactions on Information Forensics and Security.

Authors present a novel dataset for the face and facial landmarks detection in thermal images. Face detection and localization of facial landmarks are the primary steps in many face applications. Numerous algorithms and datasets have been introduced to achieve accurate detections in the wild. However, varying conditions of illumination still pose challenging problems. In this regard, thermal cameras are widely used because of their operation on longer wavelengths (without requiring the illumination or light). Thermal face and facial landmark detection in the wild is an open research problem because most of the existing datasets were collected in controlled conditions. In addition, many of them were not annotated with the face bounding boxes and facial landmarks.

ISSAI

CONFERENCES

PARTICIPATION IN CONFERENCES 2022
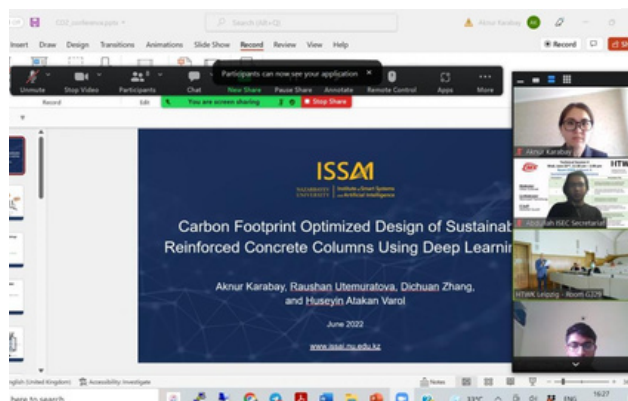
"CARBON FOOTPRINT OPTIMIZED DESIGN OF SUSTAINABLE REINFORCED CONCRETE COLUMNS USING DEEP LEARNING" RESEARCH PAPER PRESENTED AT EURO-MED-SEC-4

On 22 June, ISSAI data scientist Aknur Karabay presented a research paper entitled "Carbon Footprint Optimized Design of Sustainable Reinforced Concrete Columns Using Deep Learning" by Aknur Karabay, Raushan Utemuratova, Dichuan Zhang, Huseyin Atakan Varol at the Fourth European and Mediterranean Structural Engineering and Construction Conference (EURO-MED-SEC-4).

EURO-MED-SEC is organized annually by the International Structural Engineering and Construction Society (ISEC). EURO-MED-SEC-4 has been ISEC Society's most successful EURO-MED conference to date, with 128 papers published.

ISSAI'S DOUBLE DEBUT AT LREC 2022

During the week of 20–25 June, 2022, the 13th Edition of the Language Resources and Evaluation Conference (LREC) was held in Marseille, France.

For the first time in the history of LREC, the Institute of Smart Systems and Artificial Intelligence (ISSAI) presented two papers – KazakhTTS2: Extending The Open-Source Kazakh TTS Corpus With More Data, Speakers, And Topics by Saida Mussakhojayeva, Yerbolat Khassanov, Huseyin Atakan Varol and KazNERD: Kazakh Named Entity Recognition Dataset by Rustem Yeshpanov, Yerbolat Khassanov, Huseyin Atakan Varol. KazakhTTS2 paper was presented remotely by ISSAI postdoctoral scholar Yerbolat Khassanov, ISSAI technical writer Rustem Yeshpanov gave an on-site poster presentation on KazNERD.
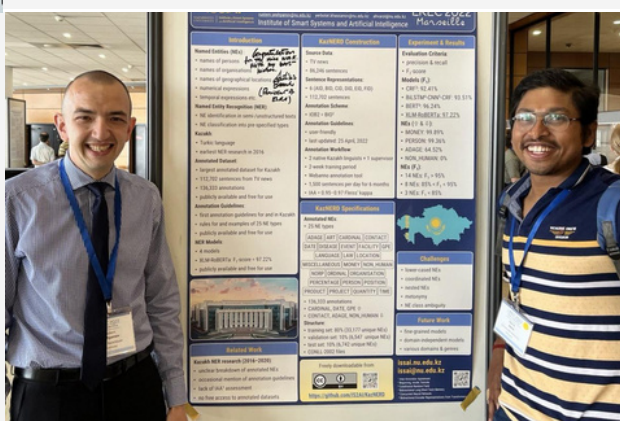
ISSAI RESEARCH PAPER PRESENTED AT ODYSSEY 2022: THE SPEAKER AND LANGUAGE RECOGNITION WORKSHOP

The research paper "A Study of Multimodal Person Verification Using Audio-Visual-Thermal Data" by Madina Abdrakhmanova, Saniya Abushakimova, Yerbolat Khassanov, and Huseyin Atakan Varol was presented by ISSAI data scientist Madina Abdrakhmanova at Odyssey 2022: The Speaker and Language Recognition Workshop.

In the paper, the authors explore an approach to multimodal person verification using audio, visual, and thermal modalities. The combination of audio and visual modalities has already been shown to be effective for robust person verification. From this perspective, the authors investigate the impact of further increasing the number of modalities by adding thermal images. The experiment conducted demonstrated the superior performance of the trimodal verification system.

ISSAI

CONFERENCES

PARTICIPATION IN CONFERENCES 2022

ISSAI'S PROJECT WINS FIRST PRIZE IN THE SCIENTIFIC VIDEO POSTER COMPETITION OF THE 8TH INTERNATIONAL SUMMIT ON MEDICAL & PUBLIC HEALTH NUTRITION EDUCATION & RESEARCH

ISSAI's project has taken first prize in the scientific video poster competition of the 8th International Summit on Medical & Public Health Nutrition Education & Research. The project "AI Applications for Dietary Interventions: Perspectives from East and Central Asia" is a joint scientific work of ISSAI and Nazarbayev University School of Medicine (NU SOM), authored by Aknur Karabay, Huseyin Atakan Varol, and Yen Mei Chan. The poster was presented by ISSAI data scientist Aknur Karabay.

The NNEdPro International Summit on Nutrition and Health is a leading annual scientific meeting. The theme of this year's summit is Empowering Global Nutrition with Digital Technology.

ISSAI RESEARCHER ZHANAT MAKHATAEVA PRESENTS HER AR PROJECT AT THE EDTECH FORUM 2022

ISSAI researcher and NU PhD student Zhanat Makhataeva presented her ExoMem project at the Fourth Annual EdTech Forum held in Nur-Sultan, June 27-29. The forum was held offline as part of the Astana Finance Days organized by the Astana International Financial Center. This year's key theme was Mobility. Edutainment. Technology. Action (M.E.T.A).

Speakers at the M.E.T.A panel session discussed human-robot interaction in education, current and future education trends and challenges. Zhanat gave a presentation on the use of AR and AI and their application in human assistance. She introduced the audience to her ExoMem project, an AR-based human memory augmentation system, and provided examples of other state-of-the-art technologies that can assist human vision, hearing and speech functions.

ISSAI PARTICIPATES IN THE CONFERENCE "AI IN DETECTING FALSIFIED FOOD PRODUCTS", ORGANIZED BY THE FAO/WHO COORDINATING COMMITTEE FOR EUROPE OF THE CODEX ALIMENTARIUS COMMISSION (CCEURO)

ISSAI data scientist Aknur Karabay and NU School of Medicine Professor Mei-Yen Chan participated in the conference "AI in detecting falsified food products", held by the Food and Agriculture Organization/World Health Organization Coordinating Committee for Europe of the Codex Alimentarius Commission (FAO/WHO CCEURO). The aim of the conference was to raise awareness of AI methods used to detect counterfeit food. These methods are crucial for the fast detection of counterfeit food products in times of global challenges, such as the global hunger crisis predicted by the UN FAO.

Aknur and Prof. Chan gave a presentation on their current research project "Potential Applications of AI for Food Authentication and Dietary Interventions".

ISSAI

CONCERENCES

CONFERENCES

PARTICIPATION IN CONFERENCES 2022

ISSAI DATA SCIENTIST SAIDA MUSSAKHOJAYEVA PRESENTS THE KAZAKH SPEECH CORPUS 2 AT INTERSPEECH 2022

ISSAI Data Scientist Saida Mussakhojayeva presented ISSAI's recent project—Kazakh Speech Corpus 2 (KSC2): An Industrial-Scale Open-Source Kazakh Speech Corpus by Saida Mussakhojayeva, Yerbolat Khassanov and Huseyin Atakan Varol—at the 23rd INTERSPEECH conference held in Incheon, Republic of Korea from September 18 to 22.

INTERSPEECH is the world's largest and most comprehensive conference on the science and technology of spoken language processing. INTERSPEECH conferences emphasize interdisciplinary approaches addressing all aspects of speech science and technology, ranging from basic theory to advanced applications. This year the conference was held in Songdo ConvensiA, in Incheon, Korea, under the theme "Human and Humanizing Speech Technology".

FROM 24 TO 28 OCTOBER 2022, NAZARBAYEV UNIVERSITY (NU) IS HOSTING THE ASIAN UNIVERSITY ALLIANCE (AUA) YOUTH FORUM. THE MAIN THEME OF THIS FORUM IS "BUILDING A META SILK ROAD: INTEGRATION OF ASIAN CULTURES IN THE REAL AND VIRTUAL UNIVERSE".

Within the framework of the forum, the Institute of Smart Systems and Artificial Intelligence (ISSAI) organized a two-hour workshop together with professors from the Departments of Computer Science (CS) and Robotics and Mechatronics of the School of Engineering and Digital Science, NU.

The main aim of the workshop was to present the progress of ongoing research in artificial intelligence, computer vision, wireless sensors, robotics, and smart systems.

ISSAI MEMBERS PARTICIPATE IN THE CONFERENCE "THE INFORMATION SPACE OF THE KAZAKH LANGUAGE AND THE EXPERIENCE OF BUILDING THE NATIONAL CORPUS", HELD BY THE INSTITUTE OF LINGUISTICS NAMED AFTER AKHMET BAITURSYNULY

On 15 November 2022, ISSAI members, Kuralay Baimenova, Saida Mussakhojayeva, and Rustem Yeshpanov participated in the conference "The Information Space of the Kazakh Language and the Experience of Building the National Corpus", organized by the Institute of Linguistics named after Akhmet Baitursynuly.

The presentation by the ISSAI researchers focused on AI projects for the Kazakh language, namely Kazakh Speech Corpus, Kazakh Speech Corpus 2, Kazakh Text-to-Speech, Kazakh Text-to-Speech 2, and Kazakh Named Entity Recognition Dataset. The presenters explained the projects and showed their demos to the conference attendees. In addition, projects such as Uzbek Speech Corpus, Turkish Speech Corpus and Turkic ASR and TTS were briefly mentioned as part of the corpus development initiatives.

ISSAI

CONFERENCES

PARTICIPATION IN CONFERENCES 2022

ISSAI RESEARCHERS PARTICIPATE IN THE HIGH PERFORMANCE COMPUTING (HPC) AND ARTIFICIAL INTELLIGENCE (AI) SYMPOSIUM

On 28 th November, ISSAI researchers participated in the High-Performance Computing (HPC) and Artificial Intelligence (AI) Symposium, organized by Nazarbayev University (NU) School of Sciences and Humanities professors Dr. Ernazar Abdikamalov, Dr. Sergey Bubin, Dr. Michael Lewis, and Dr. Bekdaulet Shukirgaliyev. The aim of the event was to raise awareness of current projects and computational resources, promote interaction, and collaboration between different research groups, and get a better picture of NU's computational needs to drive the ongoing development of HPC & AI resources at the University.

ISSAI members spoke about what computing resources they have and how they use them for their projects.

ISSAI MEMBERS PARTICIPATE IN INTERNATIONAL TELECOMMUNICATION UNION'S ITU-D STUDY GROUP MEETING HELD IN GENEVA, SWITZERLAND

On October 2022, ISSAI, as the leading academic institution in the field of AI in Central Asia, was officially invited to contribute to the ITU-D sector. The ITU-D consists of two groups (Group 1 and 2). For each group, one-week meetings were held in late November (Group 1) and early December (Group 2). Assigned to Group 2, ISSAI had the honor of presentig its achievements in promoting the Kazakh language in the digital world.

During the week, ISSAI technical project coordinator Yerbol Absalyamov and senior data scientist Askat Kuzdeuov gave a presentation on the development of the Kazakh Speech Corpus and the challenges faced in the course and gave a detailed overview of the data collection and model development processes. Yerbol and Askat introduced the audience to ISSAI's attainments in developing the Kazakh Text-to-Speech (KazakhTTS) dataset and the Kazakh Named Entity Recognition Dataset (KazNERD). They also spoke about ISSAI's research work in multilingual speech recognition for Turkic languages and provided examples of intergovernmental cooperation with the Image and Speech Processing Laboratory in the Department of Computer Systems of Tashkent University of Information Technologies, which led to the creation of the Uzbek Speech Corpus (USC).

ISSAI

PUBLICATIONS

RESEARCH PUBLICATIONS IN 2022

End-to-End Sequential Indoor Localization Using Smartphone Inertial Sensors and WiFi

BY MUKHAMET NURPEIISSOV, ASKAT KUZDEUOV, ASLAN ASSYLKHANOV, YERBOLAT KHASSANOV, HUSEYIN ATAKAN VAROL

The 2022 IEEE/SICE International Symposium on System Integration (SII 2022). 9-12 January, 2022 Norway. Online participation - Kuzdeuov A.

KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus

BY SAIDA MUSSAKHOJAYEVA, YERBOLAT KHASSANOV, HUSEYIN ATAKAN VAROL

"Interspeech 2022 18-22 September 2022, Incheon, Korea"

Deep Robust Control of a Mechatronic System with Parametric Uncertainties

BY DAULET BAIMUKASHEV, YERZHAN RZAGALIYEV, MATTEO RUBAGOTTI, HUSEYIN ATAKAN VAROL

2022 IEEE/ASME International Conference on Advanced Intelligent Mechatronics

Design and Development of Wire-Driven Pulley Guided Continuum Robot (WPCR) Arm and Friction Analysis

BY AZAMAT YESHMUKHAMETOV, KOICHI KOGANEZAWA

2022 IEEE/ASME International Conference on Advanced Intelligent Mechatronics

ISSAI

PUBLICATIONS

RESEARCH PUBLICATIONS IN 2022

An Augmented Reality-Based Warning System for Enhanced Safety in Industrial Settings

BY TOLEGEN AKHMETOV, HUSEYIN ATAKAN VAROL

IEEE Transactions on Industrial Informatics, 2022, doi: 10.1109/TII.2022.3216009

PhotoElasticFinger: Robot Tactile Fingertip Based on Photoelastic Effect

BY DINMUKHAMMED MUKASHEV, NURDAULET ZHUZBAY, AINUR KOSHKINBAYEVA, BAKHTIYAR ORAZBAYEV, ZHANAT KAPPASSOV

Sensors 2022, 22(18), 6807

SF-TL54: A Thermal Facial Landmark Dataset with Visual Pairs

BY ASKAT KUZDEUOV, DARINA KOISHIGARINA, DANA AUBAKIROVA, SANIYA ABUSHAKIMOVA, HUSEYIN ATAKAN VAROL

2022 IEEE/SICE International Symposium on System Integration (SII)

A Study of Multimodal Person Verification Using Audio-Visual-Thermal Data

BY MADINA ABDRAKHMANOVA, SANIYA ABUSHAKIMOVA, YERBOLAT KHASSANOV, HUSEYIN ATAKAN VAROL

The Speaker and Language Recognition Workshop (Odyssey 2022) 28 June - 1 July 2022, Beijing, China

PUBLICATIONS

RESEARCH PUBLICATIONS IN 2022

TFW: Annotated Thermal Faces in the Wild Dataset

BY ASKAT KUZDEUOV, DANA AUBAKIROVA, DARINA KOISHIGARINA, HUSEYIN ATAKAN VAROL

IEEE Transactions on Information Forensics and Security (Volume: 17)

Carbon Footprint Optimized Design of Sustainable Reinforced Concrete Columns Using Deep Learning

BY AKNUR KARABAY, RAUSHAN UTEMURATOVA, SICHUAN ZHANG, HUSEYIN ATAKAN VAROL

Proceedings of International Structural Engineering and Construction, 9(1), 2022 State-of-the-art Materials and Techniques in Structural Engineering and Construction

KazNERD: Kazakh Named Entity Recognition Dataset

BY RUSTEM YESHPANOV, YERBOLAT KHASSANOV, HUSEYIN ATAKAN VAROL

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 417–426 Marseille, 20-25 June 2022 © European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

KazakhTTS2: Extending the Open-Source Kazakh TTS Corpus With More Data, Speakers, and Topics

BY SAIDA MUSSAKHOJAYEVA, YERBOLAT KHASSANOV, HUSEYIN ATAKAN VAROL

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5404–5411 Marseille, 20-25 June 2022 © European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

ISSAI

OPPORTUNITIES FOR KAZAKHSTANI STUDENTS

Summer Research Program 2022

This year, the ISSAI team has again organized a summer research program for students from various Kazakhstani universities. The program offers students the opportunity to gain knowledge in artificial intelligence, machine learning, programming and related fields by working closely with ISSAI data scientists. During the program, students are assigned a mentor from the ISSAI data science team with whom they work on real-life AI projects.

The Summer Research Program is an initiative of the ISSAI team that first took place in the summer of 2021 and has been held every summer since. The main goal of the program is to give students from all over Kazakhstan the opportunity to gain hands-on experience working on AI projects, acquire knowledge and skills in the field of AI, and gain real-world experience working with data scientists. The program lasts approximately two months during the summer, when students typically have their summer vacation.

To be admitted, applicants must go through a selection process. The students with the best results and academic performance are then selected. This year, we received applications from universities, such as Eurasian National University, Kazakh British Technical University, Suleyman Demirel University, Yale University, Nazarbayev University, Skolkovo Institute of Technology, Payame Noor University, Al-Farabi Kazakh National University and others.

ISSAI

MAIN NEWS

—

The President of the Republic of Kazakhstan Kassym-Jomart Tokayev visits Nazarbayev University and the Institute of Smart Systems and Artificial Intelligence for the first time

On May 4, the President of the Republic of Kazakhstan Kassym-Jomart Tokayev visited Nazarbayev University in order to familiarize himself with the educational and research activities as well as the strategy for further development of the higher educational institution.

During the visit, NU President Shigeo Katsu welcomed the Head of State and presented a master plan for the development of the campus, research centers and laboratories of Nazarbayev University.

President Tokayev was presented with projects in the fields of science, artificial intelligence, education, healthcare, modernization of the political system, digitalization, agriculture, ecology and sports with which the university contributes to the implementation of the initiatives of the Head of State.

In particular, the Head of State familiarized himself with the projects of the Institute of Smart Systems and Artificial Intelligence (ISSAI). ISSAI Founding Director, Prof. Huseyin Atakan Varol, gave a visual presentation on ISSAI's research work and activities. Prof. Varol expounded on the Institute's projects, including AI-based speech and natural language processing technologies for the Kazakh language, AI solutions for the Ministry of Finance and advancements in healthcare, and reported on the educational initiatives for students in Kazakhstan. Following this, ISSAI Technical project coordinator, Yerbol Absalyamov, demonstrated the technologies developed by the ISSAI team for automatic Kazakh speech recognition and Kazakh text-to-speech conversion.

In addition, researchers Dr. Zhanat Kappassov, Togzhan Syrymova, and Dr. Azamat Yeshmukhametov presented their scientific projects to the President. Particularly, Dr. Kappassov and Ms. Syrymova demonstrated their project "Tactile Glove for Breast Lump Detection", and Dr. Yeshmukhametov introduced the President to the project "Tensegrity Robot: A Novel Robot Design Paradigm".

ISSAI

MAIN NEWS

━━━━

President of the 76th Session of the United Nations General Assembly, Abdulla Shahid, visits ISSAI

On 12 August 2022, the President of the 76th Session of the United Nations General Assembly, Abdulla Shahid, visited the Institute of Smart Systems and Artificial Intelligence (ISSAI) lab as part of his official visit to Nazarbayev University (NU).

NU President Shigeo Katsu welcomed the distinguished guest to campus, showed him the University's master plan, and gave him a brief tour of the campus, including a visit to the ISSAI lab.

ISSAI Founding Director Professor Atakan Varol welcomed the distinguished guest and gave him a tour of the ISSAI lab, followed by ISSAI Data Scientist Askat Kuzdeuov, who gave an introductory presentation on ISSAI projects. Afterwards, ISSAI PR Manager Kuralay Baimenova spoke about the Institute's natural language processing projects and demonstrated the Kazakh text-to-speech project developed by ISSAI researchers.

In addition, researchers Daniil Filimonov and Gourav Devappa Moger spoke about their scientific projects. Daniil explained how tensegrity structures can change the perception of mobile robots, as they can withstand heavy loads and strong collisions with their low weight. One of the most impressive examples of the use of such robots is in search and rescue operations.

Abdulla Shahid is widely known as an exceptional politician and experienced diplomat. He has long served in the government and legislative structures of the Republic of Maldives. In 2021, he was elected President of the 76th session of UN GA.

ISSAI

MAIN NEWS

Minister in the Prime Minister's Office of the Republic of Singapore, Second Minister for Education and Second Minister for Foreign Affairs Dr Mohamad Maliki Osman visit ISSAI

The Minister in the Prime Minister's Office of the Republic of Singapore, Second Minister for Education and Second Minister for Foreign Affairs Dr Mohamad Maliki Osman visited ISSAI on 11 October 2022. The visit was part of a working visit to participate in the Sixth Summit of the Conference on Interaction and Confidence Building Measures in Asia (CICA), which was held in Astana from 10 to 14 October 2022.

During the visit to ISSAI, Minister Maliki was accompanied by officials from the Ministry of Foreign Affairs of the Republic of Singapore and NU Provost Ilesanmi Adesida. ISSAI Founding Director Professor Huseyin Atakan Varol acquainted the guests with ISSAI's major projects. In addition, ISSAI members demonstrated their flagship projects, such as the hybrid mobile robot based on a tensegrity structure and the tensegrity robot—a novel paradigm for robot design.

The distinguished guests were impressed by the ISSAI projects and especially praised the KazakhTTS2 project for its high accuracy.

Singapore was participating as a guest of the current CICA chair. CICA, founded by Kazakhstan in 1992, is a forum for promoting cooperation to strengthen peace, security, and stability in Asia.
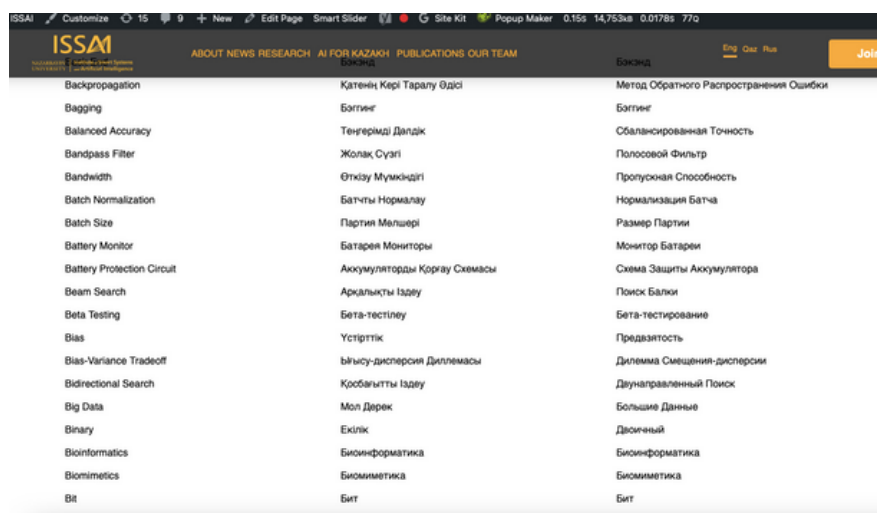
ISSAI

MAIN NEWS

On the initiative of ISSAI, the AI and robotics terminology in the Kazakh language is approved

The Kazakh translation of terms related to artificial intelligence and robotics proposed by ISSAI has been officially approved by the Language Policy Committee of the Ministry of Education and Science of the Republic of Kazakhstan. Now the Kazakh language contains translations of numerous terms and words used in the field of artificial intelligence, data analytics, robotics, and computer science.

In order to popularize the use of Kazakh in the field of data analytics and artificial intelligence, the ISSAI team developed a glossary in early 2021 containing translations of artificial intelligence terms from English into Kazakh. The glossary included more than 1000 terms. The glossary was submitted to the Language Policy Committee of the Ministry of Education and Science of the Republic of Kazakhstan for approval.

The approval process consisted of meetings of the Republican Terminology Commission with the participation of experts in Kazakh linguistics and ISSAI data scientists as experts in the field of AI and data analytics. During the discussions and voting, some words were adopted without translation and retained their original English pronunciation, but the majority received an official Kazakh translation. For example, the term artificial neural network was translated as жасанды нейрондық желі, but the term boosting was used as a loanword and translated as бустинг. In total, more than 672 words were officially approved.

The approved terms are available on termincom.kz and ISSAI's official website.

MAIN NEWS

ISSAI is sharing its AI computational resources with Kazakhstani universities

At the end of 2021, ISSAI started preparing the NVIDIA DGX-1 server to be used as a pilot server to share with the Kazakhstani universities so that they could run their research models on this server remotely for free.

Throughout 2022 ISSAI representatives held seminars on ISSAI's computational resources at various universities to promote AI computing among Kazakhstani universities. During the visits, ISSAI representatives introduced the audience to the ISSAI's mission and research activities, explained the technical features and parameters of its AI computational resources and the way of conducting research using NVIDIA DGX-1, the ability of having a remote access, and the possible exchange of researchers for collaborative research. The prospective areas of joint research are in the directions of object detection, natural language processing, face recognition, machine learning, and automated design.

ISSAI will strive to work with the other Kazakhstani universities to increase the AI research know-how and capacity of the country.
.

ISSAI

NEW PROJECTS 2022

___

Kazakh Named Entity Recognition (KazNERD)

KazNERD is the largest publicly available dataset for Kazakh named entity recognition. The dataset was built as there is a clear need for publicly available annotated corpora in Kazakh, as well as annotation guidelines containing straightforward–but rigorous–rules and examples. The dataset annotation, based on the IOB2 scheme, was carried out on television news text by two native Kazakh speakers under the supervision of the first author. The resulting dataset contains 112,702 sentences and 136,333 annotations for 25 entity classes. State-of-the-art machine learning models to automatize Kazakh named entity recognition were also built, with the best-performing model achieving an exact match F1-score of 97.22% on the test set.

The annotated dataset, guidelines, and codes used to train the models are freely available for download under the CC BY 4.0 license from
https://github.com/IS2AI/KazNERD

The link to projects description and a demo is here: https://issai.nu.edu.kz/kaznerd-eng/

NEW PROJECTS 2022

TFW: Annotated Thermal Faces in the Wild Dataset

The dataset contains thermal images acquired in a controlled indoor (c-indoor), semi-controlled indoor (s-indoor), and uncontrolled outdoor (u-outdoor) environments. The c-indoor dataset was constructed using our previously published SpeakingFaces dataset. The s-indoor and u-outdoor datasets were collected using the same FLIR T540 thermal camera with a resolution of 464×348 pixels, a wave-band of 7.5–14 μm, the field of view 24, and an iron color palette. The dataset was manually annotated with face bounding boxes and five point facial landmarks (the center of the right eye, the center of the left eye, the tip of the nose, the right outer corner of the mouth, the left outer corner of the mouth).

The link to projects description and a video is here: https://issai.nu.edu.kz/tfw-annotated-thermal-faces-in-the-wild-dataset/

ISSAI

NEW PROJECTS 2022

Kazakh Speech Corpus 2

Kazakh Speech Corpus 2 (KSC2) is the first industrial-scale open-source Kazakh speech corpus. KSC2 corpus subsumes the previously introduced two corpora: Kazakh speech corpus and Kazakh Text-To-Speech 2, and supplements additional data from other sources like tv programs, radio, senate, and podcasts. In total, KSC2 contains around 1.2k hours of high-quality transcribed data comprising over 600k utterances.

Importantly, KSC2 contains utterances with the Kazakh-Russian code-switching, a common conversation practice among Kazakh speakers.
The dataset can be used by professionals to develop various Kazakh speech and language processing applications, such as virtual assistants in the Kazakh language, robots speaking Kazakh, smart homes and cars, voice and text-enabled applications that can also assist people with special needs, and many more.
Like the first version, the KSC2 dataset is freely available to both academic researchers and industry practitioners from ISSAI website.

The link to projects description and a demo is here: https://issai.nu.edu.kz/kz-speech-corpus/

ISSAI

NEW PROJECTS 2022

▬▬▬

Kazakh language Text-to-Speech 2

In order to stimulate research and innovation and encourage the use of Kazakh in the digital field, in 2021, ISSAI developed a Kazakh speech dataset called "KazakhTTS". KazakhTTS is a high-quality open-source speech dataset that contains over 90 hours of audio recorded by professional speakers (male and female voices). The dataset has attracted great interest and has been downloaded more than 500 times in less than a year by academia and industry representatives.

To carry on the success, we present a new version of the dataset called KazakhTTS2, which includes more data, speakers, and topics. Specifically, we have increased the data size from 90 hours to 271 hours. We have added three new professional speakers (two females and one male), with over 25 hours of transcribed data for each speaker. We have diversified the topic coverage with a book and Wikipedia articles.
KazakhTTS2 dataset can be used to develop Kazakh text-to-speech models for numerous applications, such as interactive smart assistant systems, navigation systems, announcement systems and assistive technologies for the people with special needs. Like the first version, KazakhTTS2 dataset is freely available to both academic researchers and industry practitioners from ISSAI website.

To demonstrate the utility of the KazakhTTS2 dataset, ISSAI has developed a demo program for Kazakh speech synthesis. The demo supports five different voices.

The link to project description and a demo is here: https://issai.nu.edu.kz/tts2-eng/

ISSAI

NEW PROJECTS 2022

ExoMem – Human Memory Enhancement Using Artificial Intelligence

The work presents the first human memory augmentation system that can construct a synthetic spatiotemporal memory for objects in an indoor environment. We named it as ExoMem. Our solution leverages augmented reality (AR) and artificial intelligence (AI) and comprises two components:

- AR goggles that sense the environment, exchange data over a wireless network and construct a spatiotemporal memory in the AR environment.
- A computing module that performs computer vision-based localization and object detection on first-person view data received from the AR goggles.

To demonstrate the efficacy of our system, we designed object memorization and recall tasks and measured mental workload and performance during these tasks. In the memorization task, participants completed a 20-minute tour of the three floors of the building and had to memorise the location of ten different objects they saw along the path. In the recall task, participants had to recall the positions of the memorised objects in a computer-based test with and without using AR-based spatiotemporal memory.

The link to project description and videos is here: https://issai.nu.edu.kz/exomem-eng/

YOUTUBE

19551
VIEWS

712
SUBSCRIBERS

662
(HRS) WATCH TIME

Top-5
VIDEOS

- Machine Learning Basics: Ensemble Learning: Bagging, Boosting, Stacking
- Machine Learning Basics: Hidden Markov Model (HMM)
- Machine Learning Basics: Gaussian Mixture Models and the Expectation Maximization Algorithm
- A tutorial on WiFi-based Indoor Localization: Overview
- Machine Learning Basics: Independent Component Analysis

Gender

Female
34.5%

Male
65.5%

Age

13-17
18-24
25-34
35-44
45-54
54+65+

0   25   50   75   100
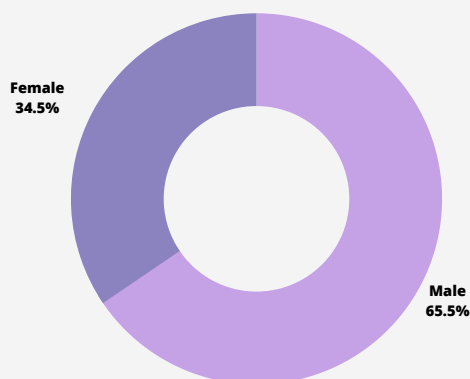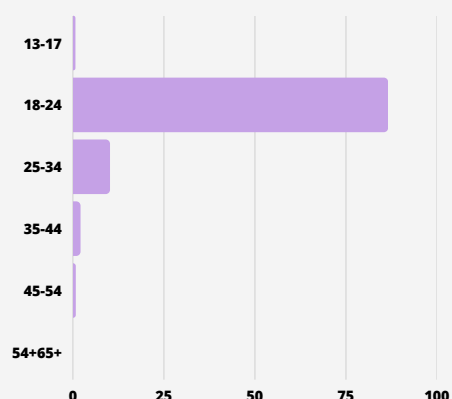
ISSAI

MEMBERSHIPS

ISSAI is a member of international organizations listed below

The International Telecommunication Union (ITU) is the specialized United Nations Agency for ICT services and technologies promotion, collaboration, and standardization.

DataCite is a leading global non-profit organization that provides persistent identifiers (DOIs) for research outputs. With the help of DataCite, organizations assign DOIs to their research works. Since August 2020, ISSAI has been a certified repository with the right to assign up to 100 DOIs.

ISSAI researchers are members of the Institute of Electrical and Electronics Engineers (IEEE), the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

NVIDIA invents the Graphic Processing Units and drives advances in AI, HPC, gaming, creative design, autonomous vehicles, and robotics. ISSAI holds 6 Nvidia DGX servers.

Overleaf is a collaborative cloud-based LaTeX editor used for writing, editing and publishing scientific documents. It partners with a wide range of scientific publishers to provide official journal LaTeX templates, and direct submission links.

ISSAI

ANNUAL REPORT 2022

**Nazarbayev University
Institute of Smart Systems
and Artificial Intelligence (ISSAI)**

Contacts

Institute of Smart Systems
and Artificial Intelligence (ISSAI)
53, Kabanbay Batyr Avenue, Block C4,
Nazarbayev University,
Astana, Kazakhstan
+7 7172 70-92-52
issai.nu.edu.kz
issai@nu.edu.kz
issai_nu