Lecture Notes in Networks and Systems 960

Fausto Pedro García Márquez
Akhtar Jamil
Alaa Ali Hameed
Isaac Segovia Ramírez *Editors*

Emerging Trends and Applications in Artificial Intelligence

Selected papers from the International Conference on Emerging Trends and Applications in Artificial Intelligence (ICETAI)

Springer

Lecture Notes in Networks and Systems 960

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the worldwide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Fausto Pedro García Márquez · Akhtar Jamil ·
Alaa Ali Hameed · Isaac Segovia Ramírez
Editors

Emerging Trends and Applications in Artificial Intelligence

Selected papers from the International Conference on Emerging Trends and Applications in Artificial Intelligence (ICETAI)

Springer

Editors
Fausto Pedro García Márquez
Ingenium Research Group
University of Castilla-La Mancha
Ciudad Real, Spain

Akhtar Jamil
National University of Computer
and Emerging Sciences
Islamabad, Pakistan

Alaa Ali Hameed
Department of Computer Engineering
Istinye University
Istanbul, Türkiye

Isaac Segovia Ramírez
Ingenium Research Group
University of Castilla-La Mancha (UCLM)
Ciudad Real, Spain

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-031-56727-8 ISBN 978-3-031-56728-5 (eBook)
https://doi.org/10.1007/978-3-031-56728-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

Preface

Emerging Trends and Applications in Artificial Intelligence

Selected papers from the International Conference on Emerging Trends and Applications in Artificial Intelligence (ICETAI)

This book is a compilation of the selected papers presented at the International Conference on Emerging Trends and Applications in Artificial Intelligence (ICETAI) in 2023.

The conference has been organized by the Istanbul Medipol University, Turkey, on September 08–09, 2023. This event brought together leading experts, researchers, scholars, and professionals from around the world to share their latest findings and explore the newest advances in the field of artificial intelligence. As technology continues to shape our lives, the role of artificial intelligence has become increasingly significant. This conference provided a unique opportunity to gain insights into the latest developments and applications of artificial intelligence in the digital age. From cutting-edge research to real-world applications, the conference provided a comprehensive overview of the field and its impact on society.

This conference managed a large number of submissions of original, high-quality research papers, where only a few were accepted. Authors submitted their work in areas related to artificial intelligence and its applications, including, but not limited to, machine learning, deep learning, computer vision, natural language processing, robotics, and more. All submissions were reviewed by a panel of experts in the field, and the accepted papers are presented in this book. This is an excellent opportunity for researchers, scholars, and professionals to showcase their work and contribute to the advancement of the field. Submissions were made through the conference website following the submission guidelines.

Each paper was peer-reviewed by at least two reviewers and evaluated based on originality, technical depth, correctness, relevance to conference, contributions, and readability. The papers were accepted based on technical merit, interest, applicability, and how well they fit a coherent and balanced technical program.

The conference was carried out in hybrid mode.

The book highlights some of the latest research advances and cutting-edge analysis of real-world case studies on computational intelligence, data analytics, IoT, and applications from a wide range of international contexts. It also identified business applications and the latest findings and innovations in Operations Management and the Decision Sciences, e.g.:

Data Analysis and Visualization

- Exploratory Data Analysis
- Statistical and mathematical modeling
- Business Intelligence

- Big Data Analysis
- Data Mining
- Cloud Computing Architecture and Systems
- ETL and Big Data Warehousing
- Business Intelligence
- Data Visualization
- Statistical Analysis

Computer Vision

- Document Analysis
- Biometrics and Pattern Recognition
- Remote Sensing & GIS
- Medical Image Processing
- Image and Video Retrieval
- Motion Analysis
- Structure from Motion
- Object Detection and Recognition
- Image Restoration
- Speech and Audio Processing
- Signal Processing

Artificial Intelligence

- Machine Learning
- Pattern Recognition
- Deep Learning
- Human–Computer Interactions
- Medical Image Processing
- Image and Video Retrieval
- Audio Video Processing
- Text Analytics
- Natural Language Processing
- Information Retrieval
- Robotics Applications

Internet of Things

- 3D Printing
- Securing IoT infrastructure
- Future of IoT and Big Data
- Internet of Things
- Intelligent Systems for IoT
- Security, Privacy, and Trust
- Visual Analytics IoT
- Data Compression for IoT Devices
- IoT Services and Applications
- Education and Learning
- Social Networks Analysis

Communication Systems and Networks

- Antennas, Propagation and RF Design
- Transmission and Communication Theory
- Wireless/Radio Access Technologies
- Optical Networks and NGN
- 5G & 6G Cellular systems and SON
- Sensor Networks
- Multimedia and New Media
- High-Speed Communication.
- Computational Intelligence in Telecommunications

Software Engineering

- Requirements Engineering
- Security Aspects
- Agile Software Engineering
- Software Evolution & Reuse
- Reverse Engineering
- Software Dependability
- Data & AI Monetization and Products
- Data as a Service/Platform
- Biomedical Experiments and Simulations
- Decision Support Systems

Fausto Pedro García Márquez
Akhtar Jamil
Alaa Ali Hameed
Isaac Segovia Ramírez

Organization

General Chairs

Gökhan Silahtaroğlu	School of Business and Management Sciences, Istanbul Medipol University, Turkiye
Samee U. Khan	Department of Electrical & Computer Engineering, Mississippi State University, USA

Advisory Committee

Ahmed Abdelgawad	Central Michigan University, USA
Gabriella Casalino	Università degli Studi di Bari Aldo Moro Bari, Italy
Hasan Dincer	Istanbul Medipol University, Turkiye
Rene Vinicio Sanchez Loja	Universidad Politécnica Salesiana, Ecuador
Serhat Yuksel	Istanbul Medipol University, Turkiye

Program Chairs

Fausto Pedro Garcia Marquez	University of Castilla–La Mancha, Spain
Kivanç Kök	Istanbul Medipol University, Turkiye

Organizing Committee

Kemal Özdemir	Istanbul Medipol University, Turkiye
Bahadır Güntürk	Istanbul Medipol University, Turkiye
Elif Baykal	Istanbul Medipol University, Turkiye
Humera Azam	University of Karachi, Pakistan
Özge Doguc	Istanbul Medipol University, Turkiye
Kailas Hambarde	University of Beira Interior, Portugal
Kevser Şahinbaş	Istanbul Medipol University, Turkiye
Tuba Khalifa	Istanbul Medipol University, Turkiye
Esra Baytören	Istanbul Medipol University, Turkiye
David Yeregui Marcos	University of Leon, Spain

Technical Program Chairs

Alaa Ali Hameed	Istinye University, Turkiye
Akhtar Jamil	National University of Computer and Emerging Sciences, Pakistan

Publication And Publicity Chairs

Dilek Yomralıoğlu	Istanbul Medipol University, Turkiye
Nada Misk	Istanbul Medipol University, Turkiye

Registration Committee

Sümeyye Özdemir	Istanbul Medipol University, Turkiye
Müge Eke	Istanbul Medipol University, Turkiye

Technical Program Committee

Abd Ullah Khan	National University of Sciences and Technology, Pakistan
Ahsan Altaf	Senior RF Systems Expert, Ahsan Altaf, Sweden
Alfredo Peinado Gonzalo	University of Castilla-La Mancha, Spain
Ali Javed	University of Engineering and Technology Taxila, Pakistan
Ali Osman Serdar Çıtak	Istanbul Medipol University, Turkiye
Alimatu-Saadia Yussiff	University of Cape Coast
Ameur Bensefia	Higher Colleges of Technology, Abu Dhabi, UAE
Atınç Yılmaz	Istanbul Beykent University, Istanbul, Türkiye
Aymen M. Khodayer Al-Dulaimi	Al-Farahidi University, Iraq
Bharat Bhushan	Sharda University, India
Chawki Djeddi	University of Rouen, France
Enkeleda Lulaj	University Haxhi Zeka, Kosovo
Faezeh Soleimani	Ball State University, USA
Ferhat Özgür Çatak	University of Stavanger, Norway
Francoise Contreras	Colombia, Universidad del Rosario
Ghulam Abid	Kinnaird College For Women, Pakistan
Haroon Rashid	Universiti Kebangsaan Malaysia/Xpertopedia Academy, Malaysia

Hasan Ali Khattak	National University of Sciences and Technology, Pakistan
Ihsan Ali	University of Malaya, Malaysia
Isaac Segovia Ramírez	University of Castilla-La Mancha, Spain
Isidro Peña García-Pardo	University of Castilla-La Mancha, Spain
J. Satpathy	Management University of Africa, Kenya and Srinivas University, India
Kiran Sood	Chitkara University, India
Mahavir Arjun Devmane.	VPPCOE & VA, India
María del Valle Fernández Moreno	University of Castilla-La Mancha, Spain
Marina Karpitskaya	Yanka Kupala University, Belarus
Momina Mustehsan	Bahria University, Pakistan
Muhammad Bilal	University of Engineering and Technology, Lahore, Pakistan
Muhammad Ilyas	Altinbas University, Turkiye
Muhammad Zeshan Alam	Brandon University, Canada
Muhammed Davud	Istanbul Sabahattin Zaim University, Turkiye
Muhsin Jaber Jweeg C	Al-Farahidi University, Iraq
Muneshwar Rajesh Niranjan	Amrutvahini College of Engineering Sangamner, India
Murat Kuzlu	Old Dominion University, Norfolk, VA, USA
Mustafa Al-asadi	KTO Karatay University, Turkey
Mustafa Takaoğlu	TÜBİTAK-BİLGEM, Kocaeli, Turkey
Naresh Kumar	University of Nizwa, Oman
Natalia Markovskaya	Yanka Kupala University, Belarus
Öznur Gülen Ertosun	Istanbul Medipol University, Turkiye
Özlem İlday	Istanbul Medipol University, Turkiye
Pedro José Bernalte Sánchez	University of Castilla-La Mancha, Spain
Rajasekaran S.	University of Technology and Applied Sciences, India
Raed Khalid Ibraheem	Al-Farahidi University, Iraq
Rana Atabay	Istanbul Medipol University, Turkiye
Reda CHEFIRA	Private University of Marrakesh, Morocco
S. G. Gollagi	KLE College of Engineering and Technology, India
Salih Sarp	Virginia Commonwealth University, Richmond, VA, USA
Şebnem Özdemir	Istanbul İstinye University, Istanbul, Türkiye
Serkan Eti	Istanbul Medipol University, Istanbul, Türkiye
Shivaji Ramdas Lahane	R. H. Sapat College of Engineering Management Studies and Research Nashik, India
Sibel Senan	Department of Computer Engineering, Istanbul University, Turkiye

Sobhan Sarkar	University of Edinburgh, UK
Subhan Ullah	National University of Computer and Emerging Sciences, Pakistan
Syed Attique Shah	Birmingham City University, UK
Tareq Hafdi Al-Khayat	Al-Farahidi University, Iraq
Ufuk Baytar	Istanbul Topkapı University, Istanbul, Türkiye
Ümit Çelebi	Istanbul Medipol University, Turkiye
Varsha Bhole	A.C. Patil College of Engineering, India
Vinayakumar Ravi	Prince Mohammad Bin Fahd University, Saudi Arabia
Waleed Ead	Beni-Suef University, Egypt
Waleed Khalid Ibraheem	Al-Farahidi University, Iraq
Yurdagül Meral	Istanbul Medipol University, Turkiye
Zeynep Orman	Istanbul University, Turkiye

Contents

Formation of a Speech Database in the Karakalpak Language for Speech Synthesis Systems

N. S. Mamatov[1(✉)], K. M. Jalelov[1], B. N. Samijonov[2], A. N. Samijonov[3], and A. D. Madaminjonov[1]

[1] Digital Technologies and Artificial Intelligence, Tashkent Institute of Irrigation and Agricultural Mechanization Engineers" National Research University, Tashkent, Uzbekistan
m_narzullo@mail.ru
[2] Sejong University, Seoul, South Korea
[3] Tashkent University of Information Technologies Named After Muhammad Al-Khwarizmi, Tashkent, Uzbekistan

Abstract. The article deals with speech synthesis, the conversion of an arbitrary text given in a natural language into spoken form in that language, and this issue has been of interest to mankind since ancient times. Also, the paper described mechanical, electrical, articulating, format, linear prediction, concatenative, selective, and statistical parametric synthesis methods. The paper describes how a high-quality speech database is essential for creating speech synthesis systems. Also, the paper notes that such databases have been created for many languages, but no speech database has been created to translate text information in the Karakalpak language. Therefore, creating a speech database for converting Karakalpak text into speech is urgent. Furthermore, this article depicts the problems and their solutions related to the creation of a speech database in the Karakalpak language.

Keywords: text-to-speech (TTS) · speech database · synthesis system · digital signal processing (DSP)

1 Introduction

The Karakalpak language, a member of the Turkic language family, holds the official status in the Republic of Karakalpakstan, an autonomous region within the Republic of Uzbekistan. About a million people speak this language, which combined with Kazakh and Nogai, two other Turkic languages that are a part of the Kipchak language group, makes up the Kipchak language family. The Khorezm, Navoi, and Bukhara regions, as well as nearby Kazakhstan and Turkmenistan, the Russian Federation, and Afghanistan, are the main distribution areas for Karakalpak language speakers. The two north-eastern and two south-western dialects of the Karakalpak language are primarily separated from one another phonetically. The literary Karakalpak language evolved during the initial decades of the 20th century, drawing inspiration from the northeastern dialect. It boasts

F. P. García Márquez et al. (Eds.): ICETAI 2023, LNNS 960, pp. 333–343, 2024.
https://doi.org/10.1007/978-3-031-56728-5_28

a remarkable history and culture, making it a fascinating linguistic development. The Karakalpak people's traditional music, dance, and literature are all expressed in this language, which is regarded as their mother tongue. Speaking of several Turkic languages, it is widely utilized and of strategic importance in Central Asia. Communication and understanding amongst the many ethnic and linguistic groups in the area are made easier by knowing and comprehending the Karakalpak language.

The Karakalpak language is also important for scientific research. Linguists and anthropologists can learn a lot about the history and culture of the Karakalpak people by studying their language. The development of a text synthesis system in the Karakalpak language is expected to contribute to the advancement of language technologies and the preservation of this language for future generations.

The creation of a text-to-speech (TTS) synthesis system in any natural language, including the Karakalpak language, requires a speech database. It is employed to instruct the system on how to produce the proper intonation, rhythm, and sounds of the language. This necessitates using a wide variety of speech, including official and informal speech as well as speech with various emotions, intonations, and accents.

The ability to produce high-quality synthetic speech will help make the language more approachable to a wider audience, including non-fluent speakers, as speech technologies advance. As a result, the Karakalpak language will be encouraged to be used in a variety of contexts, including business, entertainment, and education. This will also assist to conserve the language for future generations.

This article discusses the technical and linguistic considerations of speech data recording and the importance and sources of speech data recording for native speakers.

TTS is a cutting-edge technology that transforms written text into spoken language, utilizing algorithms and advanced techniques to analyze and comprehend the written material to produce precise speech output.

The fundamental components of TTS systems include text analysis, linguistic processing, and speech synthesis. The text analysis module scrutinizes the written content. This breaks the text down into manageable chunks, like words, phrases, and sentences. The language processing component then applies grammar, pronunciation, and intonation rules to provide a phonetic representation of the text. The phonetically represented text is then translated into speech that is sufficient for hearing through the speech synthesis component [1].

TTS systems can produce speech that sounds robotic or natural-sounding and mimics human voice and intonation. TTS technology has many different uses. These include speech recognition technology, automated voice assistants, and aids for people who have trouble speaking [2].

2 Methodology

In general, a TTS synthesizer consists of a text analysis module and a digital signal processing (DSP) module (Fig. 1). The text analysis module serves to create a phonetic transcription of the read text with the necessary intonation and rhythm. The DSP module enables the production of synthetic speech that is suitable for transcription by the text analysis module (Fig. 2).

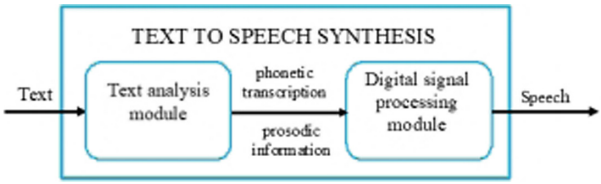

Fig. 1. General functional scheme of TTS synthesis

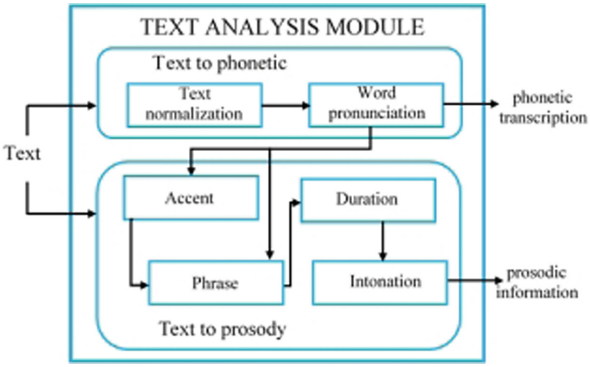

Fig. 2. TTS system text analysis module

The text analysis step is very complex because this step only needs to generate all the information required by the DSP module (for speech generation) from the text. However, plain text does not contain all the information needed to create a speech. The first block of the text analysis module converts the entered text into phonetic transcription (Text-to-Phonetics (T2P)), and the second block produces prosodic information from the text (Text-to-Prosody).

The text-to-phonetic transcription block can be divided into such modules as text normalization and word pronunciation. Below is a brief description of these two modules:

The supplied text is organized by the text normalization module using predetermined word lists. It stretches any numbers, acronyms, or idioms it finds into full text. Usually, basic grammar is used for this.

Word pronunciation The pronunciation of a string of words is established once they have been generated using the text normalization module. The simple letter-to-speech (LTS) rule is used when words are spoken exactly as they are written. When the contrary situation arises, a morphosyntactic analyzer becomes necessary. This tool is designed to group words within sentences based on their syntactic relationships, including nouns, verbs, and adjectives, while also classifying speech elements like prepositions, stems, and adverbs according to their specific properties. A lexicon is then used to establish how they should be pronounced.

Turning from text to prosody. The term prosody refers to certain features of the speech signal, such as loudness in pitch, i.e., intonation, or vocal changes in tempo, duration, stress, and rhythm. The naturalness of speech is mainly described in terms of

prosody. Prosodic events are also called suprasegmental events because these events do not coincide with segments (sounds, phonemes), but with syllables or groups of syllables.

How speech sounds are pronounced, known as prosody, plays a significant role in communicating the meaning of sentences. To generate the appropriate prosody for a given text, a text-to-prosody block is utilized. This block utilizes both the text itself and the output of the word pronunciation module to produce prosodic information. The text-to-prosody block can be further divided into several subprocesses, each of which is responsible for determining stress, phrasing, duration, and intonation for every sentence. Below is a brief description of these four processes:

The accent is based on word order. The placement of words in a sentence determines where the emphasis is placed, and this can greatly impact the way the sentence is interpreted. When constructing a sentence, it is important to consider the role of different types of words. For instance, content words such as nouns, adjectives, and verbs are usually given greater stress compared to function words such as articles, prepositions, and conjunctions. Understanding these patterns of emphasis can aid in predicting the intonation and duration of the sentence. By considering these factors, one can ensure that the intended meaning and emphasis of the sentence are accurately conveyed to the listener or reader.

Phrases are divided into phrasal verbs, and phrase boundaries are set that match the text, and these boundaries make it possible to restore pauses and intonation contours.

Intonation determines the type and meaning of a sentence. The sentence can be neutral, a command, or a question. In addition, intonation can be used to determine speaker features such as the speaker's gender, age, and emotions. The intonation module creates pitch contours for sentences. For example, "Esikti ash" and "Esikti ashıń" have different prosody. In terms of intonation contour, the first sentence is imperative and has a relatively flat pitch contour, while the second is interrogative, signifying a tone that ascends towards the sentence's conclusion.

Duration Segmental continuity is an important aspect of prosody and affects the general rhythm of speech, stress, syntactic structure of the sentence, and speech rate. Many other factors affect the length of a speech segment.

Digital signal processing (DSP) module. This module uses phonetic transcription and prosodic data generated based on the text analysis module for speech formation. This can be done in two ways, namely:

- using a set of rules describing how one phoneme affects another (called the coarticulation effect)
- storing different copies of each speech sound unit and using them as the final acoustic units.

The two primary categories of TTS systems, synthesis by rule and synthesis by combination, have been developed based on the methods described above. The DSP module's overall structure is illustrated in the accompanying figure (Fig. 3).

A TTS system is a sophisticated piece of computer software that combines several parts to create synthesized speech from written text. The following will be the TTS system's primary elements:

Text analysis. The TTS system's text analysis component analyzes the incoming text to separate and categorize individual text components including words, phrases,

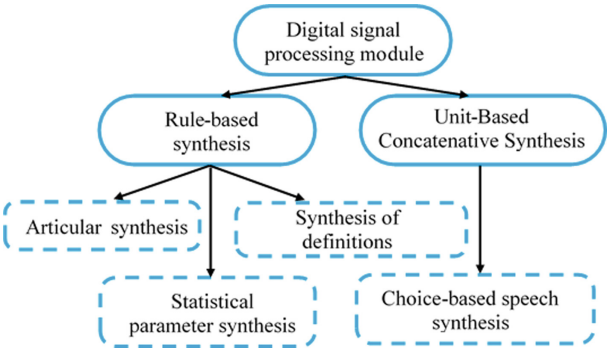

Fig. 3. Schematic of a digital signal processing (DSP) module

and sentences. To analyze the grammatical structure of the text, it carries out tasks like tokenization, assigning part-of-speech labels, and parsing [3].

Linguistic processing. The TTS system's linguistic processing section uses linguistic models and rules to translate a text's syntactic expression into a phonetic representation. Prosody, stress, and intonation, as well as pronunciation modeling, fall under this category.

Speech synthesis. Using a speech synthesis component, the TTS system converts the phonetic representation of the text into audible speech. To create understandable and natural-sounding speech sounds, it makes use of a variety of signal-processing techniques, such as filtering, amplification, and modulation (Fig. 4).

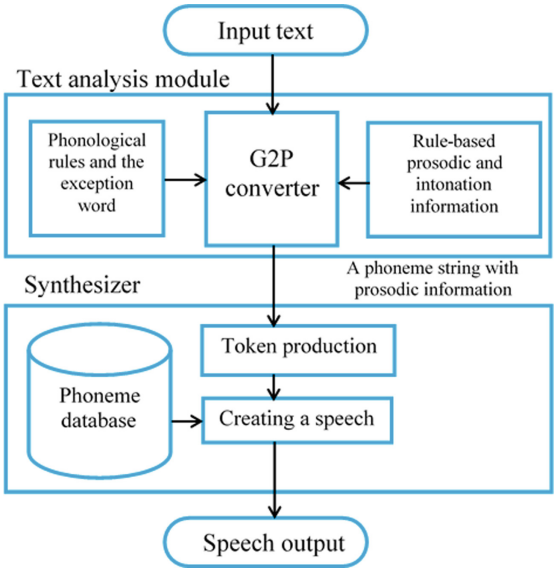

Fig. 4. Speech synthesis system

Sound database. This is a collection of speech samples written by native speakers. Samples are used to train a TTS system to produce speech that closely resembles natural speech. The database employs phonetic transcriptions of speech samples for the training of linguistic and speech synthesis models.

User interface. It is the component of a TTS system that allows users to enter text and interact with the system, and it can be a command-line interface, a web interface, or an API.

Audio Output: The audio output component of a TTS system is used to convert synthesized speech into an audio format that can be played on a loudspeaker or other audio device [4].

The database typically necessitates speech samples from a native speaker proficient in the desired language, along with phonetic transcriptions and linguistic annotations [5]. Some of the ways in which a speech database is important to a TTS system include:

Teaching the speech synthesis model. The TTS system makes extensive use of machine-learning techniques to generate natural-sounding speech. The speech database serves to train the speech synthesis model of the system, enabling it to learn and compare the written text with the corresponding speech sounds. The more high-quality speech data a system has access to, the better it will be at generating natural-sounding speech [6].

Improve pronunciation accuracy. The speech database can be used to improve the system's accuracy in pronouncing words and phrases. By providing the system with multiple examples of how a native speaker would pronounce words, it can be taught to pronounce them correctly and consistently.

Support for multiple voices. Typically, a TTS system uses speech databases to generate multiple voices, each with its features. By training the system on speech data from different speakers, it can be trained to produce different voices with different pitches, tones, and other features.

Strengthening intonation and prosody. A high-quality speech database helps the TTS system produce speech with natural-sounding intonation and prosody [7]. This speech output can be represented as a human, and its diagram is shown in the figure below (Fig. 5).

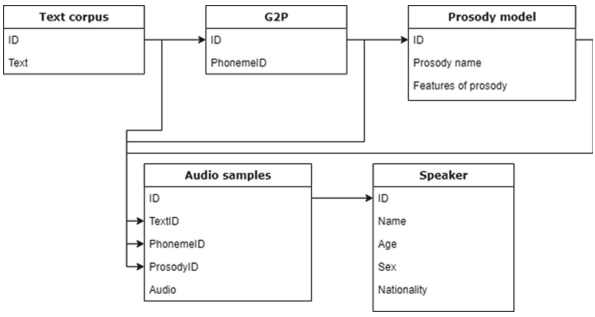

Fig. 5. The scheme of expressing speech like a human

In the absence of a speech database, a TTS system relies solely on rule-based algorithms to generate speech, resulting in less clear and less natural-sounding speech.

The creation of a speech database for the Karakalpak text-to-speech (TTS) synthesis system was carried out in the following main steps:

Choosing a speaker. When creating a speech database, it is very important to select a Karakalpak speaker, that is, a speaker when recording speech samples. It requires the speaker to have a clear, natural, and beautiful pronunciation and to read the text coherently and expressively. It is also important that the speaker have a popular and pleasant voice.

Recording speech samples. A chosen speaker is tasked with reading an extensive range of sentences in a dedicated room equipped with high-quality recording devices. The speech samples need to encompass diverse speech sounds, comprising vowels, consonants, and diphthongs, while also capturing various prosodic elements like stress, intonation, and rhythm. Speech samples should cover a variety of contexts, including different sentence types, word positions, and phonetic environments.

Transcription of speech samples. It is desirable to phonetically transcribe the recorded speech samples by a linguist. Phonetic transcriptions must represent the exact pronunciation of each speech sound in the sample.

Explanation. Speech samples must be annotated with additional linguistic information. For example, part-of-speech tags, prosodic cues, and contextual information can be interpreted. This information helps the TTS system generate more natural-sounding speech.

Initial processing. Pre-processing of recorded speech samples can also be performed to remove noise, normalize pitch, and adjust pitch and tempo. This helps to ensure that the speech samples in the database are of high quality and consistency [8–12].

TTS system training. The final step is to use the speech database to train the TTS system's speech synthesis model. Based on the patterns and features of the speech samples in the model database, it learns to match the corresponding speech sounds to the written text [7]. Then, through the TTS system, it can form high-quality, natural-sounding speech in the Karakalpak language. Building a speech database for the Karakalpak TTS synthesis system presents certain challenges. Some of these issues are outlined below.

Dialect and phonological diversity. There are several dialects of the Karakalpak language, each of which has its own phonological features. This can make it difficult to choose a single vowel that represents the entire language. In this case, it is required to record speech samples from several speakers representing different dialects and phonological features.

Limited resources for linguistic annotations. Linguistic annotations of speech samples are important in building a high-quality speech database. However, resources for this task may be limited, such as linguists or software tools.

Technical limitations. The recording equipment and software used to create a speech database must be of high quality and capture the subtle features of speech sounds. In addition, the TTS system itself may have technical limitations that affect the quality of speech output [18].

For the following reasons, it is necessary for the speaker who records the speech samples for the database to be fluent in his native language.

Naturalness. A native speaker of a language can produce speech that is natural and unique. They become relatively more familiar with important features of the language, including intonation, stress, and rhythm. This has a strong impact on the quality of speech generated by the TTS system.

Compatibility. Only a native speaker can provide coherent speech patterns that represent the language as a whole. They may also have consistent pronunciation and intonation. This allows the TTS system to provide natural-sounding speech output that is representative of the language.

Phonological accuracy. A native speaker is more likely to have a thorough understanding of the phonology of the language needed to correctly transcribe speech patterns. This TTS system serves to generate speech output that accurately reflects the pronunciation of the language.

Some of the important technical aspects of speech database recording depend on the equipment used, the recording environment, and the features of the recording technique used.

Equipment. The quality of the recording equipment used can significantly affect the quality of the speech samples obtained. High-quality microphones and recording software are required to capture high-definition recordings. For example, a condenser microphone with a flat frequency response can record a wide frequency range and create clear and detailed recordings.

Environment. The recording environment must be chosen wisely to minimize various surrounding noises and interferences. A soundproof booth or studio is usually used to achieve this. It is also desirable that the recording environment be free of echoes that may affect the quality of speech samples. For example, acoustic foam panels can be used to reduce reflections and echoes.

Techniques. Nowadays, many techniques can be used to obtain high-quality speech samples. For example, the speaker may be required to speak clearly and consistently with appropriate pauses between words and sentences [19].

Recording format. The format of the recorded speech samples also affects the quality of the speech database. Common formats used for speech databases include WAV, MP3, and FLAC. These formats have different relationships between file size and sound quality. For example, uncompressed formats such as WAV provide the highest resolution.

In addition to the technical aspects of creating a speech database for a TTS system, there are also linguistic aspects.

Speech style. To cover the speech styles used in regular conversation, a speech database must have a variety of styles, such as formal, informal, and casual. Diverse speech tempos and accents should be captured on audio.

Dialect. A speech database is also required to take into account the different dialects and accents used in natural language. To ensure the speech database encompasses various dialects, recordings can be collected from individuals residing in different regions.

Context. It is essential to gather speech samples from diverse settings such as phone conversations, public speaking, and group discussions. By incorporating a wide array of speech contexts into the database, the synthesized speech can be adapted to suit various scenarios effectively.

Feeling. A speech database must contain speech samples representing various emotions, including joy, anger, sadness, and fear. It allows natural expression to express different emotions in synthesized speech.

Dictionary. A speech database must also take into account the different vocabulary and sentence structures used in the target language. This can include technical jargon, slang, and colloquialisms. In the Karakalpak language, it is often necessary to use an explanatory dictionary. Because linguistic meanings, examples, and comments are given in explanatory dictionaries [20].

3 Results

The speakers read the texts at their natural pace and style in a quiet and closed environment, as well as in a noisy environment, and they were recorded. In this case, the speakers strictly followed the orthographic rules, the audio recordings of the female singers were sampled at 44.1 kHz and stored at 16 bits/sample. Male vocalists were recorded in a home studio, sampled at 48 kHz, and saved at 24bit/sample.

The resulting database consists of 19 h of audio material consisting of more than 34,000 segments. It took half a year to create the entire database, and the uncompressed data is more than 12 GB in size.

The database consists of information such as age, gender, work experience, and recording device of the speakers, and a Tacotron 2 system was used to demonstrate the use of the database. The subjective demonstration showed that the trained moles are suitable for practical use. For all listeners, the rating recordings were presented in the same order and one at a time. In this case, records were randomly selected at each stage (Table 1).

Table 1. Mean opinion score (MOS) results

System	Male	Female
Tacotron 2	4,15 ± 0,05	4,18 ± 0,08

Each listener was allowed to hear each audio recording only once, and the system was open to all listeners. Each recording was re-evaluated up to 10 times for female and male speakers. It was found that female speakers made more mistakes than male speakers.

4 Conclusions

The purpose of this article on building a speech database for a TTS system in Karakalpak is to demonstrate how to get around the challenges involved in building a voice database for TTS. The significance of a speech database in producing natural-sounding speech output is emphasized, and linguistic and technical factors including accommodating various speech contexts and styles and capturing distinctive phonemes are explored when recording a speech database.

Playing a pivotal role in developing a TTS synthesis system that meets the needs and expectations of the language's users and speakers would significantly aid in the preservation and promotion of the Karakalpak language.

The Karakalpak TTS synthesis system's speech database is a crucial part of it. Ensures that the output of the speech synthesizer is natural-sounding, clear, and appropriate for a range of situations. It is challenging to build a speech database for Karakalpak text-to-speech synthesis because there are few sources available and linguistic and technical considerations must be made. However, a high-quality speech database can be produced with careful planning and execution. As a result, more Karakalpak language speakers will be able to employ sophisticated synthesis systems.

There are numerous difficulties in developing a speech database for the Karakalpak text-to-speech technology. Since it contains several distinctive phonemes that are not found in other languages. Building a speech database with native speakers is crucial. Because only native speakers are capable of efficiently producing the necessary phonemes.

For the creation of a high-quality voice database, technical considerations including the choice of recording equipment, atmosphere, and data storage are equally crucial. The instruments utilized must be of the highest caliber and completely and precisely cover the linguistic quirks. An environment that is peaceful and devoid of outside noise is ideal to guarantee that speech samples are clear and useful.

Linguistic aspects such as ensuring the diversity of speech styles, contexts, dialects, and emotions are important in creating a speech database that accurately represents the Karakalpak language. It is natural and necessary to form a synthesized speech suitable for different scenarios.

References

1. Taylor, P.: Text-to-Speech Synthesis. Cambridge University Press, Cambridge (2009)
2. Nwakanma, I., Oluigbo, I., Izunna, O.: Text–to–speech synthesis (TTS). Int. J. Res. Inf. Technol. **2**(5), 154–163 (2014)
3. Onaolapo, J., Idachaba, F.E., Badejo, J.A., Odu, T., Adu, O.I.: A Simplified Overview of Text-To-Speech Synthesis (2014)
4. Oliveira, L.C., Paulo, S., Figueira, L., Mendes, C., Nunes, A., Godinho, J.: Methodologies for designing and recording speech databases for corpus based synthesis. In: International Conference on Language Resources and Evaluation (2008)
5. Chalamandaris, A., Karabetsos, S., Tsiakoulis, P., Raptis, S.: A unit selection text-to-speech synthesis system optimized for use with screen readers. IEEE Trans. Consum. Electron. **56**(3), 1890–1897 (2010). https://doi.org/10.1109/TCE.2010.5606343
6. Gahlawat, M., Malik, A., Bansal, P.: Natural speech synthesizer for blind persons using hybrid approach. Procedia Comput. Sci. **41**, 83–88 (2014). ISSN 1877–0509, https://doi.org/10.1016/j.procs.2014.11.088
7. Nagy, P., Németh, G.: Improving HMM speech synthesis of interrogative sentences by pitch track transformations. Speech Commun. **82**, 97–112 (2016). C (September 2016), https://doi.org/10.1016/j.specom.2016.06.005
8. Mamatov, N.S., Niyozmatova, N.A., Abdullaev, S.S., Samijonov, A.N., Erejepov, K.K.: Speech recognition based on transformer neural networks. In: 2021 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, pp. 1–5 (2021). https://doi.org/10.1109/ICISCT52966.2021.9670093

9. Mamatov, N., Niyozmatova, N., Samijonov, A.: Software for preprocessing voice signals. Int. J. Appl. Sci. Eng. **18**, 2020163 (2021). https://doi.org/10.6703/IJASE.202103_18(1).006
10. Narzillo, M., Abdurashid, S., Parakhat, N., Nilufar, N.: Automatic speaker identification by voice based on vector quantization method. Int. J. Innov. Technol. Explor. Eng. **8**(10), 2443–2445 (2019). https://doi.org/10.35940/ijitee.J9523.0881019
11. Wiedecke, B., Narzillo, M., Payazov, M., Abdurashid, S.: Acoustic signal analysis and identification. Int. J. Innov. Technol. Explor. Eng. **8**(10), 2440–2442 (2019). https://doi.org/10.35940/ijitee.J9522.0881019
12. Narzillo, M., Abdurashid, S., Parakhat, N., Nilufar, N.: Karakalpak speech recognition with CMU sphinx. Int. J. Innov. Technol. Explor. Eng. **8**(10), 2446–2448 (2019). https://doi.org/10.35940/ijitee.J9524.0881019
13. Niyozmatova, N.A., Mamatov, N.S., Tulyaganova, Sh. A., Samijonov, A.N., Samijonov, B. N: Methods for determining speech activity of Uzbek speech in recognition systems. In: AIP Conference Proceedings, vol. 2789(1), p. 050019, 23 June 2023. https://doi.org/10.1063/5.0145438
14. Mamatov, N.S., Niyozmatova, N.A., Yuldoshev, Y.S., Abdullaev, S.S., Samijonov, A.N.: Automatic speech recognition on the neutral network based on attention mechanism. In: Zaynidinov, H., Singh, M., Tiwary, U.S., Singh, D. (eds.) Intelligent Human Computer Interaction. IHCI 2022. Lecture Notes in Computer Science, vol. 13741, pp. 100–108. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-27199-1_11
15. Mamatov, N.S., Niyozmatova, N.A., Samijonov, A.N., Samijonov, B.N.: Construction of language models for Uzbek language. In: 2022 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, pp. 1–4 (2022). https://doi.org/10.1109/ICISCT55600.2022.10146788
16. Niyozmatova, N.A., Mamatov, N.S., Otaxonova, B.I., Samijonov, A.N., Erejepov, K.K.: Classification based on decision trees and neural networks. In: 2021 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, pp. 01–04 (2021). https://doi.org/10.1109/ICISCT52966.2021.9670345
17. Babomuradov Ozod, O.J., Mamatov, N.S., Boboev, L.B., Otaxonova, B.: Text documents classification in Uzbek language. Int. J. Recent Technol. Eng. **8**(2 Special Issue 11), 3787–3789(2019). https://doi.org/10.35940/ijrte.B1493.0982S1119
18. Saratxaga, I., Navas, E., Hernáez, I., Luengo, I.: Designing and recording an emotional speech database for corpus based synthesis in Basque. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy. European Language Resources Association (ELRA) (2006)
19. Iida, A., Campbell, N.: Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders. Int. J. Speech Technol. **6**, 379–392 (2003). https://doi.org/10.1023/A:1025761017833
20. Esemuratova, R., Kalendarova, M., Japaqova, R.: Qaraqalpaq tiliniń túsindirme sózligi. Qaraqalpaqstan baspası. Nókis (1982)