Ministry of Science and Higher Education of the Republic of Kazakhstan
SDU University

Ainur Kursabayeva

Study of the transformation of Kazakh language speech into text data

THESIS
Presented in Partial Fulfilment for the

Degree of Master of Technical Science in Computer Science
(degree code: 7M06012)
Department of Computer Science
Faculty of Engineering and Natural Sciences

Supervisor: **Utebayeva Dana**
Kaskelen, June 2024

SDU University
Faculty of Engineering and Natural Sciences
Department of Computer Science

Dean of Faculty of Engineering and Natural Sciences

Assistant Professor, PhD Akhmedov Ramis

_____

« 04 » 06 2024

Topic of the thesis:

Study of the transformation of Kazakh language speech into text data

Thesis submitted as part of the requirements for the award of the MSc in "7M06102 - Computer Science", SDU University

Head of Department Zhanar Mukash

Academic Supervisor Dana Utebayeva

Master student Ainur Kursabayeva

Kaskelen, 2024

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Ainur Kursabayeva

June 2024

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Dana Utebayeva, for her invaluable help and support in the process of writing my master's thesis. Her competence, attentiveness and patience had a significant impact on the results of my work.

Abstract

The transformation of speech into text data is a key component in the development of modern language technologies and artificial intelligence. Despite significant advances in this field, support for languages with unique grammatical and phonetic characteristics, such as Kazakh, remains a challenge. The purpose of this study is to analyze the existing method of converting speech in the Kazakh language into text and evaluate their effectiveness. The research methodology includes the analysis of the VOSK model for speech transformation in the Kazakh language. An experimental study is being conducted based on the KazakhTTS dataset using machine learning and natural language processing methods. The results of the experiment, presented as an indicator of the error rate in the word (WER), showed that VOSK big and VOSK small have almost the same indicators (51% and 53% respectively). It was also noted that there are limitations in recognizing word endings and that some errors occur during speech recognition. The discussion of the results highlights the potential of the model and points to the need for further improvement and training in working with more diverse data. In conclusion, the key conclusions are outlined, as well as potential directions for further research in the field of Kazakh speech recognition.

Аңдатпа

Сөйлеуді мәтіндік деректерге айналдыру заманауи тілдік технологиялар мен жасанды интеллектті дамытудың негізгі құрамдас бөлігі болып табылады. Осы саладағы елеулі жетістіктерге қарамастан, қазақ тілі сияқты бірегей грамматикалық және фонетикалық ерекшеліктері бар тілдерді қолдау күрделі мәселе болып қала береді. Бұл зерттеудің мақсаты-қазақ тіліндегі сөйлеуді мәтінге айналдырудың қолданыстағы әдісін талдау және олардың тиімділігін бағалау. Зерттеу әдістемесі қазақ тіліндегі сөйлеуді түрлендірудің VOSK моделін талдауды қамтиды. KazakhTTS деректер базасы негізінде машиналық оқыту мен табиғи тілді өңдеу әдістерін қолдана отырып эксперименттік зерттеу жүргізілуде. Word error rate (WER) бағдарламасындағы қателік көрсеткішінің көрсеткіші ретінде ұсынылған эксперимент нәтижелері VOSK big және VOSK small көрсеткіштері бірдей дерлік (тиісінше 51% және 53%) екенін көрсетті. Сондай-ақ, сөз жалғауларын тануда шектеулер бар екендігі және сөйлеуді тану кезінде кейбір қателіктер орын алатындығы айтылды. Нәтижелерді талқылау модельдің әлеуетін көрсетеді және одан әрі жетілдіру және әртүрлі деректермен жұмыс істеуге үйрету қажеттілігін көрсетеді. Қорытындылай келе, негізгі тұжырымдар, сондай-ақ қазақ тілін тану саласындағы әрі қарайғы зерттеулердің әлеуетті бағыттары көрсетілген.

Аннотация

Преобразование речи в текстовые данные является ключевым компонентом в развитии современных языковых технологий и искусственного интеллекта. Несмотря на значительные достижения в этой области, поддержка языков с уникальными грамматическими и фонетическими характеристиками, таких как казахский, остается сложной задачей. Целью данного исследования является анализ существующих методов преобразования речи на казахском языке в текст и оценка их эффективности. Методология исследования включает анализ модели VOSK для преобразования речи на казахском языке. Проводится экспериментальное исследование на основе набора данных KazakhTTS с использованием методов машинного обучения и обработки естественного языка. Результаты эксперимента, представленные в виде показателя частоты ошибок в слове (WER), показали, что VOSK big и VOSK small имеют практически одинаковые показатели (51% и 53% соответственно). Было также отмечено, что существуют ограничения в распознавании окончаний слов и что при распознавании речи возникают некоторые ошибки. Обсуждение результатов подчеркивает потенциал модели и указывает на необходимость дальнейшего совершенствования и обучения работе с более разнообразными данными. В заключение изложены ключевые выводы, а также потенциальные направления дальнейших исследований в области распознавания казахской речи.

Abbreviations

AI	– Artificial Intelligence
ASR	– Automatic Speech Recognition
GMM	– Gaussian Mixture Model
Kaldi	– Open-source toolkit for speech recognition
HMM	– Hidden Markov Model
MFCC	– Mel-Frequency Cepstral Coefficients
ML	– Machine Learning
NLP	– Natural Language Processing
STT	– Speech-to-Text
TTS	– Text-to-Speech
VOSK	– Open-source ASR toolkit
WER	– Word Error Rate

Table of Contents

Declaration i

Acknowledgements ii

Abstract iii

List of Abbreviations vi

1 Introduction 1
- 1.1 Relevance . . . 1
- 1.2 Aim, objectives and research questions . . . 3
- 1.3 The structure of the dissertation research. . . . 5

2 Literature review 6
- 2.1 Type of speech . . . 6
  - 2.1.1 Pure speech . . . 7
  - 2.1.2 Speech with background noise . . . 7
  - 2.1.3 Speech with an accent . . . 7
  - 2.1.4 Fast speech . . . 7
  - 2.1.5 Unstructured speech . . . 7
  - 2.1.6 Speech with unclear pronunciation . . . 8
  - 2.1.7 Emotional speech . . . 8
- 2.2 Difficulties with ASR . . . 8
  - 2.2.1 Human understanding of speech . . . 8
  - 2.2.2 Differences between spoken and written speech . . . 9
  - 2.2.3 Noise . . . 9
  - 2.2.4 Dialect . . . 9
  - 2.2.5 The style of speech . . . 10
- 2.3 Basic speech recognition model . . . 10
  - 2.3.1 Signal processing . . . 10
  - 2.3.2 Feature extraction . . . 11
  - 2.3.3 Modeling . . . 11
  - 2.3.4 Decoding . . . 12
- 2.4 Related work . . . 12

3 Methodology 32
- 3.1 Dataset . . . 33

3.2 Model Description 34

4 Results **37**
4.1 Settings for experimentation 37
4.2 Word Error Rate 37
4.3 Code execution time 38
4.4 Hardware resources 38

5 Discussion **39**

6 Conclusions and future work **40**
6.1 Conclusions 40
6.2 Future work 40

Bibliography **41**

Chapter 1

Introduction

In the modern world, where information flows are constantly growing and the importance of quick access to data is incredibly great, the translation of audio into text becomes of great importance. This process makes it possible to effectively convert oral speech contained in audio and video materials into a convenient and accessible text format. [1] With the development of modern artificial intelligence technologies, automatic transcription of oral speech into text has become an integral part of our digital life. This is crucial for a variety of applications, ranging from speech recognition systems on devices, ending with the creation of transcriptions for videos and audio recordings. [2] In the process of this transformation of speech into text, language features and dialects can pose serious challenges for algorithms and systems.

To comprehend the background and current status of Kazakh speech recognition technologies, it is essential to thoroughly analyze the available literature. In recent times, substantial endeavors have been dedicated to developing and enhancing speech recognition models for diverse languages. Notably, advanced deep learning models like DeepSpeech and WaveNet have demonstrated remarkable outcomes for English and other widely spoken languages.

Research into the identification of Kazakh speech is currently limited. Significant efforts in this area involve the exploration of hidden Markov models and recurrent neural networks. While these models have shown some effectiveness, they have also exposed deficiencies in accurately recognizing Kazakh dialects and accents.

In recent times, the VOSK open source speech recognition system has captured the interest of researchers owing to its versatility and capacity to function across various languages. The VOSK big and small models were developed to offer speech recognition solutions with differing levels of intricacy and data capacity. Nevertheless, their performance in relation to the Kazakh language has not been thoroughly examined yet.

1.1 Relevance

Studying the conversion of Kazakh spoken language into textual data is of great importance in today's world, where the processing of large amounts of data and

the development of artificial intelligence are becoming increasingly important. This process has a number of key aspects and perspectives that make it relevant and important for our society and the future [3].

Firstly, such research is of great importance in the context of the development of language technologies and artificial intelligence. Converting spoken speech to text is fundamental to the development of automatic speech recognition (ASR), machine translation, and other language applications [4]. These technologies are of great importance for modern communications, education and business.The development of automatic speech recognition systems in the Kazakh language makes it possible to create innovative educational applications and platforms that can provide educational content in their native language, improving the accessibility of education for Kazakh-speaking students.[5] For example, the development of applications for automatic subtitling of video lectures or audiobooks in Kazakh can make educational materials more accessible and understandable for students. Improving voice control systems in the Kazakh language can significantly improve the usability of smart home devices or mobile applications for Kazakh-speaking users. For example, creating a voice assistant that understands and responds in Kazakh can significantly improve user interaction with devices and services, especially for those who prefer to use their native language in everyday communications.Examples include various voice assistants and speech recognition technologies such as Apple's Siri, Google Assistant from Google and Alice from Yandex. When a user accesses a voice assistant in Kazakh, his oral speech is automatically transformed into a text format for analysis and processing. At the moment, these technologies do not support speech in the Kazakh language. For example, when a user requests "Alice, show the news in Kazakh", the speech recognition system translates the oral command into a text format and processes it, Alice's algorithms analyze and process the request, and then display the results of the news feed in Kazakh. Thus, the transformation of speech into text allows users to interact with Alice in Kazakh, expanding her functionality and making the application more accessible to the Kazakh-speaking audience.

Secondly, the study of this process is of great importance for the preservation and development of cultural and linguistic heritage. The Kazakh language is an important part of Kazakh culture, and its preservation and promotion in the digital world play an important role. The conversion of spoken language into text makes it possible to create digital archives, educational materials and other resources in Kazakh.

The Kazakh language has its own grammatical structure, which differs from English and other European languages. This includes the declension and conjugation of nouns and verbs, as well as specific word order rules. Properly recognizing and interpreting the grammatical features of Kazakh is a challenge for speech processing technologies. Also, the Kazakh language uses the Cyrillic alphabet, which includes several specific letters and graphemes. This includes letters such as "ə, і , ң ,ғ , ү, ұ ,қ , ө , h" that are not present in the English alphabet. This can lead to text distortion or incorrect word recognition, which complicates the correct processing of the Kazakh language in computer programs.

Studying this procedure also helps to ensure that the cultural and linguistic

aspects of the Kazakh language are preserved and passed on to future generations. The development of educational resources and digital archives in the Kazakh language helps to preserve the rich cultural legacy and sense of national identity [6]. Furthermore, the accurate identification of the grammatical constructions and distinctive symbols of the Kazakh language aids in the advancement of more precise and effective speech processing technologies, which is useful for the creation of locally tailored software and applications for Kazakh-speaking audiences.

It is important to remember that the majority of contemporary translators and voice recognition software(Yandex Translate, Google Translate) cannot recognize speech in Kazakh. This is because, as a language with little usage, Kazakh may have grammatical and phonetic elements that are peculiar for algorithms trained on more widely spoken languages. Furthermore, methods for recognising languages intended for Latin or other alphabets may face further challenges due to the Cyrillic alphabet used in Kazakh. For those who are native speakers of Kazakh, this limits access to material in the language and hinders communication.The widespread introduction of voice input reduces the relevance of developing the skill of high-speed manual typing by young specialists [7].

Thirdly, this study has practical applications in education, health care, law enforcement and many other fields. Textual data from spoken language can be used to create transcriptions of lectures, medical records, court transcripts, and many other documents, which facilitates access to information and improves processes in these fields.

The transformation of speech into text is of high importance in the context of the development of speech recognition systems, machine translation, audio indexing and digital assistants [8]. This process makes it possible to efficiently analyze and process huge amounts of audio and video data, which plays a critical role in the development of highly efficient and accurate speech recognition and interpretation systems. Moreover, the transformation of speech into text makes it accessible to a wide range of users, including people with hearing impairments, and also improves the accessibility of information and communication in the digital environment. It is important to note that this field of research is constantly evolving, requiring constant improvement of algorithms and methods for processing speech data.

All these complexities make the task of transforming Kazakh speech into textual data a challenge for researchers and engineers in the field of speech processing and artificial intelligence. Solving these problems requires specialized methods and resources, as well as consideration of cultural and linguistic contexts to achieve accurate and efficient recognition of Kazakh speech.

1.2 Aim, objectives and research questions

The primary ***aim*** of this thesis is to explore and develop an method for transforming spoken Kazakh language into textual data, with a focus on ensuring accuracy and usability.

The following aspects are included in the master's thesis tasks:

- **The study of the linguistic characteristics of the Kazakh languag**e: Conduct a thorough examination of the Kazakh language's linguistic ele-

ments, including phonetic, grammatical, and syntactic aspects that may impact the accuracy and quality of voice translation into text.

- **Existing approaches and algorithms are reviewed**: To examine existing approaches and algorithms for converting oral speech into text in various languages. To assess their efficacy and relevance to the Kazakh language.
- **Speech recognition system comparison**: Examine the various speech recognition systems on the market to assess their ability to properly convert Kazakh spoken speech into text. Compare the outcomes of different systems.
- **Evaluate the accuracy and reliability of various methods and algorithms**: develop a methodology and criteria for assessing the quality of translation of speech in the Kazakh language into a text format.

Research questions:

1. What is the accuracy and efficiency of existing methods for transforming Kazakh speech into text data, and how do they cope with the phonetic and grammatical features of the Kazakh language?
2. What use cases are most in demand for Kazakh speech-to-text technology, and what are the benefits and limitations associated with these scenarios?

The scientific novelty of the dissertation lies in the study of the application of the voice recognition module in the Kazakh language using VOSK. It is an advanced technology developed for multilingual voice recognition [9]. As part of the research on speech recognition for this language, a new study of the VOSK big and VOSK small models in the Kazakh language using KazakhTTS data is also presented. The results and conclusions obtained complement our understanding of the potential applications and limitations of voice recognition technology in the Kazakh language, creating new opportunities for research and advancement in this field.

The theoretical significance of this work is to expand our understanding of the possibilities and limitations of speech recognition models in the Kazakh language using VOSK. The study describes in detail the process of applying modern speech recognition technologies in a specific language, which contributes to the development of a theoretical basis in the field of speech processing. The results and conclusions of the work can be used in further research to improve the accuracy and effectiveness of speech recognition in the Kazakh language, as well as in the development of new algorithms and approaches in this area.

The practical significance of this study lies in its applicability to the development and improvement of automatic speech recognition systems in the Kazakh language. The results and conclusions of the study can be used by software developers to create more accurate and effective speech interfaces, voice recognition systems, as well as in other areas where work with the Kazakh language is required. The practical application of the research results can contribute to improving the quality and accessibility of speech recognition technologies for Kazakh-speaking users, as well as contribute to the development of innovative solutions in the field of information technology.

1.3 The structure of the dissertation research.

The literature review explores different aspects and challenges of automatic speech recognition. It covers a range of speech types, such as clear speech, noisy environments, accented speech, old recordings, unstructured input, unclear pronunciation, and emotionally expressive speech. The review also addresses the complexities related to phonetic and morphological features of the Kazakh language and evaluates current methods for signal processing, feature extraction, speech modeling, and decoding. Furthermore,,it includes a thorough critique of prior research on Kazakh speech recognition to highlight key areas for future study and enhancement

The research methodology includes an integrated approach to evaluating and comparing the performance of VOSK big and VOSK small speech recognition models for the Kazakh language based on the KazakhTTS dataset. The results of this study will help identify areas for further improvement of models and the development of more accurate and reliable recognition systems for Kazakh speech.

The Result section details the parameters of the trials, such as accessing an audio file, setting up the speech recognizer, and related elements. It presents an evaluation of the VOSK big and VOSK small models based on the word error coefficient. The VOSK small model exhibited a WER of 53%, whereas the larger VOSK model displayed 51%, denoting a marginally improved accuracy with the bigger model. Additionally, it delves into code execution timing and hardware resource utilization.

The Discussion highlights the need for further improvement of the VOSK big and VOSK small models, including improving their adaptation to the specific features of the Kazakh language and improving speech recognition algorithms. It is also important to carefully select and prepare training data so that the models are more resistant to various variations of user speech.

Chapter 2

Literature review

Speech recognition has a wide range of potential applications, including voice dialing, call routing, smart home device management, keyword search, and use in communication platforms [10]. One of the particularly interesting applications is its use for monitoring and solving speech problems in children to improve the child's health [11]. Speech recognition technology allows you to control various devices without the help of hands and continues to improve accuracy by enabling emotion recognition and physiological identification of the user. In addition, it can be used to convert text to speech to help people with visual impairments. However, this aspect is not specifically addressed in this article. Developing effective speech recognition code involves challenges such as managing vocabulary within models to reduce processing time and reduce complexity as we strive for natural spontaneous speech generation. The program should also effectively recognize multiple simultaneous speech streams along with background noise. Communication platforms as a service are cloud-based service delivery models that allow organizations to integrate real-time communication services into their business applications using APIs, which simplifies access management by allowing you to use one account for each employee instead of multiple logins to various APIs. The introduction of transcription capabilities at meetings and conferences would be beneficial by improving the archiving of information using temporary labeling, as well as by increasing accessibility for users with disabilities, among other benefits. This critical feature must be integrated into communication platforms in order to remain competitive compared to competitors in the industry, while providing up-to-date functionality.

2.1 Type of speech

When analyzing the conversion of spoken Kazakh language into written text, it is crucial to take into account the different forms of speech found in the language.Speech can be divided into the following categories.

2.1.1 Pure speech

Pure speech is a standard form of speech reproduction characterized by clarity, accuracy and the absence of external acoustic interference or distortion.[12] This type of speech provides optimal conditions for speech recognition, since acoustic signals are presented clearly and consistently. The main characteristics of clear speech are precise pronunciation, in which phonemes and words are pronounced clearly and with the correct intonation. In addition, there is no background noise, which allows speech recognition systems to focus on the main audio signal without additional interference.

2.1.2 Speech with background noise

Speech with background sound occurs when the speech signal is accompanied by different external acoustic disturbances. These disturbances can consist of noise from vehicles, radios, televisions, conversations, or other sources that produce extra sound waves in the recording. With background noise present, acoustic signals become harder to discern and differentiate, which poses additional challenges for speech recognition systems [13]. The background noise has the ability to cover up the primary speech signal, leading to decreased clarity and interpretability for recognition algorithms.

2.1.3 Speech with an accent

Speech patterns that deviate from standard pronunciation are influenced by linguistic characteristics such as intonation, articulation, and enunciation [14]. These variations can stem from regional, national, or language-specific differences, posing challenges for speech recognition technology. They may impact the phonetic composition of words, the melody of sentences, and the speed of communication. Accommodating these linguistic traits is crucial for ensuring effective performance of speech recognition systems.

2.1.4 Fast speech

Rapid speech is defined by a fast rate of enunciating words and phrases, leading to word fusion and reduced pauses [15]. This kind of pronunciation can pose challenges for speech recognition systems as the acoustic signals become less distinct. At high speeds, speech recognition algorithms may struggle to differentiate individual sounds and words, resulting in errors when transcribing speech data into text. Quick speaking requires greater processing speed and adaptability from speech recognition systems to accurately convert spoken signals into coherent text.

2.1.5 Unstructured speech

Unstructured speaking is distinguished by its absence of a precise grammatical framework, sequence, and logical flow, presenting added difficulties for speech recognition technology [16]. This form of speech often involves spoken language

alongside question and answer patterns, as well as the inclusion of slang, idiomatic expressions, and other non-conventional linguistic elements. The diversity and deviation from standard structure in this type of communication can pose challenges for automated speech recognition systems when attempting to accurately perceive and comprehend it. Challenges may stem from ambiguities, incompleteness or variations in word pronunciation, phrases, and sentences.

2.1.6 Speech with unclear pronunciation

Speech with unclear pronunciation is defined by incomplete, inaccurate diction, or articulation issues in the speaking of words and phrases. This type of pronunciation can involve inadequate sound production, blending of words, or alteration of intonation, resulting in less clear and distinguishable speech signals. Speech recognition systems face added difficulties in processing indistinct pronunciations due to the unpredictable and diverse nature of acoustic signals [17]. These articulatory characteristics may confuse recognition algorithms and result in errors when converting spoken language into text.

2.1.7 Emotional speech

Emotive speech varies from regular pronunciation because of changes in intonation and timing that express the speaker's emotional condition. Emotions like happiness, anger, sorrow, or enthusiasm can influence vocal pitch, speaking speed, as well as breaks and emphasis in speech delivery. Recognizing emotional speech poses a specific difficulty for speech recognition systems due to the potential impact of emotions on the acoustic features of spoken language, leading to reduced predictability and stability. These alterations may involve modified intonation patterns, faster or slower speaking rates, along with adjusted stress and pauses.

2.2 Difficulties with ASR

Automatic Speech Recognition is a complex task that can face a number of difficulties in various conditions and scenarios.

2.2.1 Human understanding of speech

Emotional communication varies from regular speech patterns as it involves fluctuations in intonation and timing which mirror the speaker's emotional condition. Feelings like happiness, anger, sorrow, or enthusiasm can influence the pitch of voice, pace of speaking, and also interruptions and stresses within the spoken text [18]. This form of expression poses a specific difficulty for systems designed to recognize speech due to its potential to modify the acoustic features of language making it less foreseeable and consistent. These alterations may encompass modified intonation, faster or slower speaking speeds, along with amended emphasis and pauses.

2.2.2 Differences between spoken and written speech

In spoken Kazakh, the language often differs from its written form in several important ways. In oral communication, people tend to use more informal Kazakh spoken language often varies from the written form in several significant ways. In oral communication, individuals tend to use more casual and common words and phrases, which contributes to a sense of immediacy and authenticity in conversations. The grammar used in spoken language may also be less rigid and simplified because the focus is on conveying messages quickly and efficiently rather than strictly adhering to grammatical rules [19]. Stylistically, spoken language is more flexible and adaptable, incorporating emotional expressions, intonation variations, and informal terms that reflect its conversational nature. On the other hand, written language necessitates greater attention to word choice, grammatical structure precision, and presentation style as it aims to preserve information over time without relying on oral context for clarification. In the context of everyday oral communication among Kazakh-speaking individuals, one can observe peculiarities in the use of abbreviated forms and affective constructions. When analyzing telephone conversations in research, it was revealed that in a significant part of dialogues, reaching up to 70%, simplified forms of phrasology in the Kazakh language are used. For example, instead of the full phrase " ", which translates as "I'll call", the abbreviated form "" is often used, preserving the main meaning and semantic load of the original phrase. Similarly, instead of the full phrase " ", which translates as "Call me", a more concise construction " " can be used.Although there is no such expression in literary Kazakh.

2.2.3 Noise

In the realm of acoustic signal processing and speech recognition, noise is an unwanted element that can greatly impact the precision and quality of recognition. It originates from diverse sources like background sounds, electrical disturbances, ambient noise, and other acoustic interferences. Within speech recognition procedures, noise can alter the acoustic properties of the signal leading to challenges in accurately recognizing and interpreting speech data. Several methods and technologies for reducing noise are employed to mitigate its effects on speech recognition accuracy and quality.

2.2.4 Dialect

The transformation of dialect speech into text data is a difficult task due to the unique linguistic features that differ from the standard language or its normative forms [20]. Dialects may include differences in pronunciation, vocabulary, grammar, and phraseology, which may create additional difficulties for speech recognition systems. For example, the pronunciation of dialect words may differ from the standard one, which may lead to errors in recognizing words or phrases. Unique words and expressions specific to the dialect may be unfamiliar to speech recognition systems trained in standard forms of the language, which can also lead to errors in the transformation of speech into text. These dialect features can make it

difficult for systems to correctly recognize and interpret speech data, which makes the task of converting dialect speech into text more difficult and requires additional efforts to adapt and optimize speech recognition algorithms.

2.2.5 The style of speech

The speed of speech can range from fast to slow, reflecting a person's usual pace of speaking and information transmission[21]. Intonation and tone are used to convey emotional state, attitude towards the content, and cultural communication norms. For instance, intonation can signal confidence, insecurity, sarcasm, or friendliness. The formality of communication style is also significant in different contexts and social settings. A formal style entails using complete grammatical structures, appropriate vocabulary usage, and avoiding colloquial expressions and jargon. Conversely , an informal style tends to be more relaxed and straightforward with simpler language . Additionally , communication styles may vary based on interlocutors' relationship dynamics , cultural norms ,and societal customs.This means that some cultures promote proactive,direct communications while others encourage respectful,careful interactions.

2.3 Basic speech recognition model

An integral speech recognition model comprises various essential elements for converting a speech audio signal into written text. These elements encompass signal processing, extracting features, creating models, and decoding.

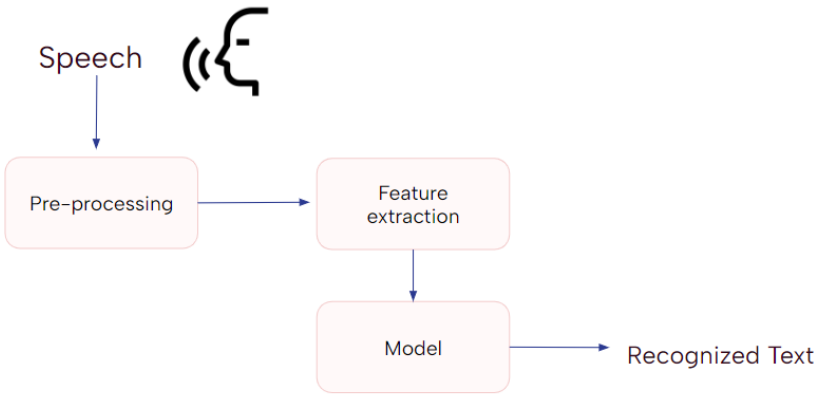

Figure 2.1 – Speech recognition procedure

2.3.1 Signal processing

Signal processing plays a crucial role in speech recognition, serving as a foundational step before delving into detailed audio signal analysis and interpretation [22]. The primary goal at this stage is to optimize and refine the audio data for

subsequent analysis. One essential early step involves filtering out noise, which has the potential to distort the audio signal and create challenges for speech recognition. Implementing specialized noise reduction and filtering algorithms can enhance signal quality by minimizing external noise distortion impact. Normalizing the volume level of the audio signal is another key process. This standardization across different recording locations ensures stability of the signals, ultimately enhancing speech recognition accuracy [23]. Additionally, dividing the audio signal into frames is an integral part of this process with each frame typically lasting 20 to 30 milliseconds - tailored for further feature extraction and analysis within speech technology applications. Subsequently after framing, preprocessing methods are applied along with feature normalization techniques such as zero alignment and mean/standard deviation standardization in order to facilitate subsequent analysis related to speech features extractions thus ensuring high performance under various acoustic conditions while preparing efficient and accurate output towards effective task completion.

2.3.2 Feature extraction

MFCC feature extraction stands as a widely utilized and efficient method for processing audio signals in speech recognition. This technique enables the depiction of the spectral features of a speech signal through a series of coefficients, considering the nuances of sound perception by humans [13]. The process commences with converting the audio signal into frequency space using Fourier transform to acquire a spectral representation for each signal frame. Subsequently, mel scaling is implemented to modify the spectral components based on a logarithmic frequency scale aligned with human ear sensitivity to different frequencies. To accommodate non-linear aspects of sound perception and mitigate spectrum dynamic range, logarithmic amplitudes are employed. The final phase involves implementing discrete cosine transform to convert the log spectrum into frequency space while retaining key components and reducing data dimensionality. The resulting MFCC coefficients undergo normalization to standardize data and enhance resilience against acoustic variations and recording conditions[24].In essence, MFCC offers an efficient portrayal of speech signals while preserving fundamental acoustic characteristics such as formants and pronunciation features. Due to this quality, MFCC proves particularly advantageous for robust speech recognition across diverse conditions and environments.

2.3.3 Modeling

In the context of speech recognition, modeling involves creating statistical models that can analyze and interpret extracted features from a speech signal. The primary objective of this process is to train the model to recognize and categorize speech data using provided training data. Speech recognition commonly utilizes various types of statistical models such as hidden Markov models, Gaussian mixture models, and deep neural networks. HMMs are probabilistic models capable of predicting likely word or phrase sequences by modeling sequences of speech features

[25]. GMMs are used to represent the distribution of speech features in feature space, often working alongside HMMs. [26] Deep neural networks have gained popularity due to advancements in machine learning and deep learning technologies, offering an effective approach for modeling speech data.[27]DNNs can be trained with large datasets to extract complex hierarchical speech features, leading to high accuracy in recognizing speeches across different conditions. For specific applications in speech recognition, diverse models may be adapted or combined depending on the task requirements and available resources. It is essential when conducting modeling activities to ensure that training data accurately represents real-world operating conditions while maintaining generalizing ability for reliable operation.

2.3.4 Decoding

Deciphering in the context of speech recognition involves interpreting and transforming identified features of a speech signal extracted from an audio file into a string of words or phrases. This process, which constitutes the final stage in the speech recognition system, encompasses several essential elements [**?**]. A primary approach to deciphering involves utilizing statistical models like hidden Markov models or deep neural networks, which have undergone training using provided data for speech recognition [28]. These models scrutinize the sequence of extracted speech features and ascertain the most plausible word or phrase sequence based on trained statistical parameters and a language model . To enhance accuracy and effectiveness during decoding, diverse strategies and techniques are frequently employed. Examples include amalgamating information from multiple sources (e.g., acoustic model, language model, dictionary) as well as implementing post-processing to rectify and optimize recognized text. Depending on the specific system and task within speech recognition, decoding may also encompass adapting the model to suit a particular speaker or recording conditions. This adaptation serves to bolster recognition quality while ensuring more precise and dependable conversion of the spoken signal into textual representation.Overall, decoding plays a pivotal role in enabling accurate and efficient transformation of spoken audio signals into comprehensible written text format through its contribution to various practical applications.

2.4 Related work

Speech recognition technology has been advancing for the past fifty years. Conventional speech recognition systems rely on a GMM-HMM approach. This technology has undergone significant developments in recent years, with a growing emphasis on addressing the challenges posed by the variability in spoken language.

Speech Recognition Evolution. The development of machines capable of reproducing human behavior, especially in the field of speech recognition and synthesis, has been attracting engineers and scientists for many years. Since Homer Dudley introduced a system model for speech analysis and synthesis at Bell Laboratories in the 1930s, significant advances have been made in automatic speech recognition technology [29]. The development of ASR technology has led to a transition

from elementary systems that recognized a limited set of sounds to complex systems capable of understanding and responding to natural speech. Breakthroughs in statistical modeling in the 1980s played an important role in significantly expanding the capabilities of ASR. Today, ASR systems are integral components of human-machine interfaces widely used in tasks such as automated call processing in telephone networks and query-based information systems. These systems make it easier to perform various functions, including providing real-time travel information, stock quotes, weather forecasts, and others.

A model-based approach to speech recognition. In the article, Hallet and Stevens aim to explore the progress in speech recognition technology, focusing on a model and program developed for research purposes.[30] The methodology used in this study includes an analysis-synthesis procedure to extract a sequence of phonemes from a temporarily varying spectrum of the input signal. The process includes comparing the input spectrum with signals synthesized according to generative rules, using a control component to transmit instructions to generative rules, and analyzing the results of preliminary signal analysis and previous attempts. In addition, the model integrates peripheral speech generation structures for processing both speech recognition and production. The main argument of this article is that by including peripheral speech generation structures, the model can effectively manage both speech recognition and speech production tasks. The recognition process is based on the calculation of phonetic parameters using generative rules, which eliminates the need for direct activation of speech structures. This approach improves the efficiency and accuracy of speech recognition systems. Overall, the article provides valuable insights in the field of speech recognition, offering a complete overview of recent developments and presenting a model and program developed for further research in this area.

Speech recognition and variability. Benzegiba et al. presented an overview of significant advances in speech processing and recognition systems, including automatic speech recognition (ARR). The study highlights important advances in this area, but also highlights the technological challenges that researchers face.[31] Special attention is paid to problems related to environmental influences, such as background noise, which can significantly complicate the speech recognition process. In addition, there is a limitation in the transmission of grammatical and semantic information. In relation to specific applications, such as a reference information service with an extensive active dictionary, as well as in the case of foreign accents, the authors emphasize the need to develop methods that can take into account the real variability of speech. As a result of the research, promising directions for the development of more efficient APP systems capable of more accurately and reliably recognizing and processing various speech variants are being considered. The study highlights the complex relationship between sociolinguistics, environment, gender, speaking pace, accent, and style, as well as regional variations and speaker characteristics. The modeling method is made significantly more complicated by these factors, especially when training resources are limited. The authors examine techniques aimed at optimizing ASR modeling and analysis, with a focus on strengthening speech variability resistance within the system.

Multi-level processing. Morgan [3] used discriminatively trained feedback

networks to review a study in the field of speech recognition. The main aim of the review was to shed light on studies that used multi-level processing before decoding word sequences using a hidden Markov model. In order to provide either a significant advantage for speech with a high signal-to-noise ratio (SNR) in large vocabulary tasks, or a significant advantage for noisy speech in small vocabulary tasks, various methods involving multiple levels of computation were described throughout the article. In the end, this review concluded that although deep processing structures could improve this genre, other criteria such as layer width and feature selection may also be important. The Neha Chadha emphasize the need to develop a recognition system that facilitates effective communication between individuals and machines through text or speech processing. They outline various speech recognition techniques and methods currently in use, along with their respective advantages and limitations. The review suggests that room for innovation exists in proposing novel approaches to the recognition process, with the potential to yield superior results compared to existing methodologies.The challenges identified include issues related to environmental noise, which degrades the performance of audio input systems. Accuracy and reliability are further compromised by unwanted input and suboptimal output results. Fault tolerance is lacking in the existing systems, and user responsiveness can be hindered when users initiate commands before resources are ready, leading to synchronization problems with multiple applications.

Tackling challenges and innovating solutions. Neha Chadha et al. have examined the current problems in the field of speech recognition and suggested possible solutions for future developments [32]. The article discusses how a recognition system is developed with the aim of improving human-machine communication based on text or speech processing. The authors' thorough analysis of the literature was undertaken with the intention of finding future research topics and solving current problems in the field of speech recognition. The authors have also described the various recognition techniques and approaches that are prevalent today, along with their respective advantages and disadvantages. The paper also emphasises the need for innovative methods for the recognition process that outperform the established ones in terms of results. It also examines the detrimental effects of external noise on the performance of audio input systems, with particular emphasis on how these effects affect accuracy, reliability and fault-tolerance. User performance issues are also addressed, particularly when dealing with unprepared resources and command synchronisation across multiple applications.

Speech technology advances. Alexandre Trilla [33] examines the complex relationship between natural language processing (NLP) and two critical areas of speech technology: text-to-speech synthesis (TTS) and automatic speech recognition (ASR). Through a comprehensive review of recent achievements, the article reveals the key role of NLP in shaping the evolution of these fields. In the field of TTS synthesis, NLP acts as a cornerstone, facilitating the conversion of input text into natural-sounding speech through careful processing of linguistic nuances. The author also emphasises the complex relationship between the preceding text processing modules and the naturalness of the resulting speech. Conversely, in ASR, NLP acts as a catalyst for simplification, optimising the recognition task

by applying pre-defined grammatical rules. This symbiotic relationship enhances the capabilities of ASR and contributes to the creation of more intuitive interfaces based on linguistic knowledge. The article discusses in detail various approaches to adapting language models, including N-gram models and the integration of context-free grammars (CFGs), for the effective use of linguistic knowledge. It also discusses innovative strategies for processing spontaneous speech using automatic generalisation, information retrieval (IR) tasks and a dialogue system. The merging of NLP and IR is also considered, emphasising the role of ontologies in building knowledge bases for advanced reasoning and understanding. Combining theoretical ideas with practical applications, this article offers a holistic view of the transformative potential of NLP in changing the landscape of language technologies.

Self-study with quantization of random projections. Chung-Cheng Chiu and et al. present a simple and effective method for improving autonomous speech recognition. The method uses a model to anticipate hidden speech signals, which are represented as individual labels generated by a random projection quantizer. The pre-training process masks speech signals and trains the model to predict labels for the masked parts. This approach shows similar results to existing advanced results on LibriSpeech with non-streaming models and is superior to wav2vec 2.0 and w2vBERT on LibriSpeech with streaming models and on multilingual tasks with non-streaming models [34]. Further analysis showed that the random projection quantizer is effective for self-learning, despite the fact that it provides a poorer representation compared to the trained VQ-VAE quantizer.The algorithm presented in the article separates the quantizer from the speech recognition model and also eliminates the need for representation training. This simpler framework makes it easier to find a good recipe for a target task. The improvement on streaming models shows that separating the quantizer from the model makes the algorithm more efficient on architectures that may be less efficient at learning representations. The improvement on multilingual tasks shows that complex tasks may benefit more from a simpler framework where finding a good recipe becomes more difficult. The analysis of quantization quality implies that representation learning is not mandatory for self-learning.

Improving speech recognition with LSTM-RNN. A study by Jane Oruch et al. presented an improved deep learning model based on recurrent neural networks (RNN) with long-term short-term memory (LSTM) to overcome this limitation [35]. In the proposed model, the RNN is integrated into a memory block as a "forgetting gate", which allows cell states to be reset at the beginning of subsequences. This adaptation allows the system to efficiently handle continuous input streams without significantly increasing bandwidth requirements. The standard architecture of the LSTM network was modified in the proposed model to make optimal use of the model parameters. The study also included the application of convolutional neural network (CNN)-based models and sequential models on the same dataset, and these models were compared with the proposed model.The proposed model avoids the problem of processing continuous input streams that are not divided into subsequences. This means that streams that are theoretically not subdivided into smaller units are easily handled by the network. The results

showed that the LSTM-RNN model outperformed other deep learning models, achieving 99.36% accuracy on a widely recognised publicly available dataset of spoken English digits.The proposed system incorporates the RNN as a forgetting gate within the network, allowing the states of cells to be reset at the beginning of subsequences. This improvement makes efficient use of network parameters and solves the computational efficiency problems associated with large networks in the context of large vocabulary speech recognition.

Deep Learning in Multilingual ASR. Amodei et al. demonstrated the potential of using an integrated deep learning method for speech recognition in English and Mandarin Chinese, two languages with unique characteristics [36]. By replacing traditional manual development pipelines with neural networks, this approach provides increased adaptability to work with various speech scenarios such as noisy environments, accents, and multilingual contexts. The most important element of their approach is the use of high-performance computing technologies, which led to a noticeable 7-fold acceleration compared to their previous system [**?**]. This increased efficiency allows for faster experiments and iterations to determine optimal architectures and algorithms.

CNN Integration in ASR Architecture. Study [37] have shown that hybrid architectures based on a deep neural network and a hidden Markov model are effective in improving speech recognition compared to traditional systems based on a Gaussian mixed model and HMM. The improvement is largely due to DNN's ability to capture complex correlations between speech features. To further reduce the error rate, this study examines the integration of convolutional neural networks into the existing DNN-HMM structure. The authors present an overview of the principles of CNN and their applicability in speech recognition, as well as a specialized scheme with limited weight distribution designed to improve the presentation of speech characteristics. Unique architectural aspects such as local connectivity, weight distribution, and pooling allow CNNS to account for differences arising from different speakers and environmental conditions, providing a certain level of invariance to minor changes in speech characteristics along the frequency axis. The experimental results demonstrate a significant reduction in the error rate when integrating CNN into the speech recognition pipeline, achieving improvements in the range from 6% to 10% compared to traditional DNNs in both TIMIT phone recognition tasks and voice search tasks for a large vocabulary.

Signal Processing in ASR. The guide [38] provides a comprehensive overview of the signal processing techniques used in modern speech recognition systems, with an emphasis on the most commonly used methodologies. During the discussion, four main operations underlying signal modeling are considered: formation of spectral characteristics, analysis of spectral components, parameter transformation and statistical modeling.The tutorial describes three important achievements that have occurred in the field of speech recognition over the past five years. Firstly, a variety of parameter sets that combine absolute spectral data with dynamic or time-dependent spectral information are becoming increasingly common. Secondly, similarity transformation methods for effective normalization and decorrelation of parameters have become widespread. Finally, the problem of estimating signal parameters has become an integral part of the speech recognition process,

which makes it possible to use more advanced statistical models to estimate the signal spectrum in a closed loop.

Improved speech recognition. Anupam Choudhary and Ravi Kshirsagar describe in detail the process of speech recognition using artificial intelligence techniques. It includes an acoustic model, a language model, a trigram model, a class model, and a source and channel model. Speech recognition or natural language processing refers to artificial intelligence methods for communicating with a computer in a natural language such as English [39]. The purpose of NLP (Natural Language Processing) is to understand IP and take action. The article notes that external factors such as noise, environmental conditions and device placement can significantly affect speech recognition performance. Vocabulary is highlighted as a critical factor affecting productivity. The construction of a language model and vocabulary corpus tailored to a specific application is emphasised. The use of a personalised dictionary for specific applications is recommended. The article mentions several factors that affect speech recognition performance, including the quality and location of microphones and speakers, emotional and physical states, speech rate, voice quality, and the size and shape of the vocal tract. The choice of speech model is discussed, with the recommendation to use a general speech model for everyday spoken English, but to choose a limited model when recognition is intended for a specific application. A restricted model configured for the application can improve accuracy and speed.

NLP advanced integration. Katalin Ungurean and Dragos Burileanu [40] give an overview of the NLP stage in the Romanian language TTS system and describe its integration into the Speech Synthesis Markup Language (SSML) system, which is currently the widely accepted standard for TTS document authorship and intermodule exchange.To obtain a modular approach to the SSML stack, they separated the application task from the speech generation task, using SSML to exchange data and control both internally between different NLP modules and externally between the aforementioned tasks. In the model, the application is based on the NLP algorithms described above. Each individual NLP module will progressively add value to the text or SSML string provided as input in order to obtain an appropriate phonetic and prosodic representation of the original text message. This approach can offer more flexibility because both the NLP and the speech generation engine can be independently replaced or upgraded at any time, or even different levels of NLP can be easily replaced.The article describes the main elements of the advanced NLP structure for the Romanian speech synthesis system TTS. The results are mostly obtained statistically, but rules or even hybrid approaches are also used. An important feature that has not been highlighted is that our NLP approaches, in addition to achieving very good results, are essentially text independent. In other words, they successfully solve the problem of lexical stress positioning, syllable division and letter-sound conversion for any invisible word that conforms to Romanian lexicology and phonology rules.

Analysis and prospects of speech synthesis in the text. Thierry Dutoit's article [4] is a comprehensive introduction to the current state of the art in speech-to-text synthesis, focusing on the digital signal processing (DSP) and natural language processing (NLP) components. It examines the potential ap-

plications and challenges of text-to-speech systems, emphasising their importance in telecommunications services, education, assisting people with disabilities, multimedia and other areas. It also emphasises the importance of understanding synthetic utterances for these applications, while naturalness is not always the main problem. He also created several sub-modules in the OEA module, such as pre-processing, morphological analysis, contextual analysis and syntactic-prosodic analysis. The complexity of phonetic transcriptions is also discussed, especially in cases of homographs and pronunciation ambiguities, through the Letter-Sound Conversion module. A distinction is also made between rule-based synthesis and concatenation synthesis.

Improve ASR.Alexander Trilla's article is devoted to the application of natural language processing (NLP) methods in two important areas of speech technology: text-to-speech synthesis (TTS) and automatic speech recognition (ASR)[33]. The article provides an overview of the latest achievements in the field of NLP in various disciplines.The study highlights the critical importance of NLP in analyzing input text for voice conversion in the context of TTS synthesis. The effectiveness of previous text processing modules strongly affects the quality of the final speech. NLP improves the recognition process in ASR by assuming that incoming speech follows established grammatical rules. According to the article, NLP can improve the capabilities of ASR by providing more natural interfaces with a certain degree of language training. It examines N-gram language models and contextual cues as ways to update language models to use linguistic information. In addition, the study suggests approaches to working with spontaneous speech using automated generalization, as well as to assist in information retrieval and dialog systems. It examines the importance of ontologies for building knowledge bases that support logical reasoning.

CNN and LSTM. Daniel S. Park and et al. are introducing special add-ons that directly expand the capabilities of ASR neural networks [41]. This method, in particular, is used to filter bank coefficients.First, the authors describe their approach to ASR tasks using the Listen, Attend and Spell (LAS) networks. LAS networks are comprehensive models that are well suited for ASR and have established standards in this area. They describe in detail the network architecture consisting of a two-layer convolutional neural network (CNN) with maximum integration, a bidirectional encoder based on LSTM and a two-layer RNN decoder. The input data are logarithmic spectrograms, and the output data is a series of attention vectors for generating decryption markers.The authors' work has produced impressive results, especially with regard to the Libre Speech 960h and Switchboard 300h datasets. In LibriSpeech, they achieved a word error rate (WER) of 6.8% when tested without a language model and 5.8% when combining surface with a language model. These results are comparable to the performance of previous modern hybrid systems. For the Switchboard dataset, they achieved 7.2% in the Hub5'00 test suite for the switch/CallHome task without using a language model and 6.8% when using shallow fusion with a language model. These results are also consistent with the performance of previous modern hybrid systems. The improvement, along with LAS networks and embedded learning methods, represents a significant advance in ASR technology, pushing the boundaries of accuracy

and performance in speech recognition tasks.

End-to-end Learning. In a piece authored by Avni Hannun and et al.[42] A novel end-to-end deep learning method for voice recognition is showcased. This novel technique is a more straightforward substitute for conventional speech recognition systems, which sometimes need intricate processing steps and have trouble operating in loud settings. Using these novel strategies, the Deep voice system outperforms previously reported findings in the speech recognition space. He scored an astounding 16 points based on the findings of the complete test set of closely examined Switchboard Hub5'00 data.One of the system's most noteworthy accomplishments is that it can function effectively in challenging and loud circumstances, outperforming popularly utilized sophisticated commercial speech recognition systems in comparable settings. This technology is more adaptable to a broad range of applications, from virtual assistants to transcribing services, and is also more effective due to the elimination of hand-crafted components and concentration on direct learning. The remarkable accomplishments of Deep Voice show how it might impact voice recognition technology in the future.

Attention-Based Processes. The adaptation and improvement of attention-based processes for voice recognition tasks is investigated by Jan Chorowski et al. To enhance an attention-based recurrent sequence generator's (ARSG) phoneme identification performance, the authors develop and enhance several strategies. By using these techniques, the model's resilience is intended to be increased and difficulties caused by variable-length input sequences are addressed.In order to recognize phonemes on the TIMIT dataset, the authors first modify a machine translation model [43]. They are able to obtain a competitive phoneme error rate (PER) of 18.7%. They do note, though, that this modified model has issues with input sequences that are shorter than the ones it was trained on. Due to this constraint, creative methods for improving the attention mechanism are developed. According to the test results, performance is improved both by using windowed mode and by smoothing. The study showed a relative increase of 3.7% compared to the base model when using convolutional functions, and when using survey methods, this performance increases by 5.9%. Interestingly, short-term repetitions did not cause confusion in the basic model, as it learned to coordinate effectively. The paper argues that these new attention mechanisms, such as improved normalization and feature extraction, could potentially be used in more extensive applications than speech recognition, such as neural Turing machines and picture caption creation.

Comparative Performance. In an article written by Yogita H. Gadage and Sushama D. Shelke, presented a multilingual speech-to-text conversion system focused on converting speech signals into text representations in several languages[44]. The system operates in two stages: training and testing, and uses a combination of methods, including MFCC feature extraction, minimum distance classifier, and SVM for classification.The system uses a combination of minimum distance classifier and SVM to classify words. A word with a minimal difference between its features and the reference feature vectors is recognized as output data.The proposed system provides promising accuracy for multiple languages. It is noteworthy that for the Marathi language, the system achieves an accuracy of 93.625%

using MFCC feature extraction, minimum distance classifier and SVM combination. This performance surpasses the accuracy achieved using the MFCC feature extraction method and the CDHMM classifier, which resulted in an accuracy of 88.80% for Marathi. The system also provides 91.6667% accuracy for English and 90.625% for mixed language scripts in English and Marathi.

Speech-to-Text Translation. Yu-An Chung and et al. presented the concept of developing speech-to-text (ST) translation systems [45] that do not rely on parallel bilingual corpora. Instead, they use monolingual sets of speech and text data. The paper describes a structure that allows to build speech-to-text translation systems without the need to use parallel bilingual corpora. This approach solves the problem of uncontrolled translation, making it more accessible and universal.The platform first creates an ST system that can perform basic word-by-word translation. The article presents the results of experiments using their uncontrolled ST method. They compared their approach with controlled baselines by performing several training runs. The results show that their VecMap system outperformed MUSE in several experiments. This demonstrates the effectiveness of VecMap, especially in complex scenarios with weak, completely uncontrolled initialisation.

ASR System Improvement. Karita et al. We have improved the speed and accuracy of the ASR system by combining Transformer with ACER's RNN-based achievements[46]. The model included a connection-based time classification for co-learning and decoding with Transformer, which led to faster learning and improved LM integration. This proposed ASR system has demonstrated significant improvements in solving various tasks, reducing WER from 11.1% to 4.5% for The Wall Street Journal and from 16.1% to 11.6% for TED-LIUM due to the integration of CTC and LM into the basic version of Transformer.

Dynamic probability adjustment. Yangyang Shi and his colleagues [47] presented a technique for suppressing weak attention, which dynamically adjusts the probability of concentration in order to reduce unnecessary focus on excessive acoustic frames. This approach prioritizes suppressing attention to past frames rather than future ones, and has demonstrated a reduction in the frequency of word errors compared to traditional transformer models. According to the results of Libre Speech testing, the WAS method resulted in a 10% decrease in WER for pure testing and a 5% improvement for streaming transformers, which brought it to an advanced level among streaming models.

RNN to Transformer Conversion. The RNN-based encoder-decoder model in the E2E system has been replaced by the Transformer architecture in Chang X's ans et al. work[48], in order to use this model in a multichannel beamforming network, the self-attention component was modified so that it was limited to a segment rather than the entire sequence, reducing computational complexity. In addition to the architectural improvements, preprocessing was also included to predict weighted errors of external reverberation, which allowed the model to process reflected signals. Experiments with the expanded wsj1-2 mixed housing show that transformer-based models achieve better results in the absence of echo in single-channel and multi-channel modes, respectively.

Continuous Russian Speech Recognition. Researchers from Russia conducted a study on continuous Russian speech recognition using DNN confidence

[49]. They used a method in which finite state machine-based converters were used for speech recognition. The study showed that their proposed approach led to an increase in the accuracy of speech recognition compared to hidden Markov models. In addition, the study compared language models created using direct and recurrent neural networks. Three different implementations were considered: LIMSI software tools for building a direct-acting neural network with a limited output level; a direct-acting neural network with clustering using the entire dictionary; and a recurrent neural network with clustering. Experimental results have shown that language models built using a direct-coupled neural network work less efficiently than models built using recurrent neural networks, which, in particular, leads to an increase in efficiency by 0.4% when using a recurrent network compared with direct communication on test data[50].

Neural networks play a key role in reducing the dimensionality of features for phone recognition systems. Hongbing Hu and Stephen A. Zakarian propose two approaches using neural networks within a hidden Markov model to achieve this goal [51]. By training neural networks as feature classifiers, they effectively diminish the number of dimensions while enhancing differences between speech features. The research findings indicate that the transformed features yield slightly improved recognition accuracy compared to the original ones, showing significant superiority over linear dimensionality reduction methods. Notably, with an extensive number of training iterations on TIMIT data, phone recognition achieves a maximum accuracy of 74.9

Multilingual DNN. The study [52] presents a new method of multilingual speech recognition using a multilingual DNN with a common hidden level. The SHL-MDNN architecture uses common hidden layers for multiple languages, while the softmax layers remain language-dependent. Experimental results show that this approach can reduce the number of errors by 3-5% compared to monolingual DNNs trained only on the basis of language data. In addition, it shows that common hidden layers studied in different languages can effectively improve the accuracy of recognizing new languages, while reducing the number of errors from 6% to 28%. It is important to note that these improvements are observed even for target languages belonging to other language families than those used to explore common hidden layers. In conclusion, the SHL-MDNN approach shows promising results in improving the efficiency and accuracy of multilingual speech recognition systems by using common feature transformations in various language contexts.

Kazakh Speech recognition Several studies have introduced speech databases for the Kazakh language.Makhambetov and colleagues created a Kazakh language corpus with approximately 40 hours of transcribed read speech data obtained in a sound-proof recording studio [53]. Similarly, Mamyrbayev et al. [54] gathered 123 hours of data using professional recording facilities, which were later expanded by Mamyrbayev et al.. Khomitsevich et al [55]. used 147 hours of bilingual Kazakh-Russian speech data for the development of code-switching ASR systems. Shi et al. [56] published 78 hours of transcribed Kazakh speech recordings from 96 Chinese students, while the IARPA Babel project provided a Kazakh language pack comprising about 50 hours of conversational and 14 hours of scripted telephone speech; however, these databases are unfortunately either not publicly accessible

or lack sufficient size for robust Kazak ASR system development.

Voice activity detection system [57] based on a multitasking learning U-shaped neural network, demonstrating high performance compared to existing models. The training and testing of the model were carried out using the TIMIT speech corpus for English. In addition, a multilayer neural network was trained and tested to determine speech activity based on the same TIMIT speech corpus. However, further research is needed using datasets from different languages and analyzing the required number of native speakers for an effective VAD system.

DNN is widely used in speech research for speech recognition, and it has been proven to produce positive results. For example, study [58] presented a system for recognizing spontaneous Czech, Slovak and Russian speech in interviews with Holocaust witnesses. Automated basic transcriptions based on certain rules have been created, and many transcription options have been developed to account for phonetic phenomena such as consonant assimilation at word boundaries. In addition, the study described the variants of spoken pronunciation of the Russian language and differences in accents among residents of different countries, including Ukraine, Israel and the United States. The corpus used to develop acoustic models of the Russian language consisted of 100 hours of data and included DNN. A bigram model using the Katz scheme with an offset was used as a language model; it included a dictionary volume of 79 thousand transcriptions, which led to incorrect word recognition in 38.57% of cases.

Despite being classified as a language with limited resources, extensive research has been carried out on the enhancement of Kazakh speech recognition systems for applications including voice-controlled devices, speech-to-text conversion, and automatic translation.

Khassanov and et al. [59] presented an open source Corpus of Kazakh Speech, which contains about 332 hours of transcribed audio recordings. It covers more than 153,000 statements from participants from different regions, age groups and genders. KSC has been carefully evaluated by native speakers of the Kazakh language to ensure its high quality. This publicly available database is the largest of its kind and is designed to support various applications for processing Kazakh speech and language learning. The authors provide comprehensive information about the procedures for collecting and preprocessing data, the technical characteristics of the database, as well as their experiences and problems they encountered during the creation process. Initial speech recognition experiments conducted using KSC showed promising results: the error rate in characters was 2.8%, and in words - 8.7%. In addition, in order to increase the reproducibility of the experiment and facilitate the use of the corpus, they also released the Asp network recipe for the speech recognition model.

The open-source Kazakh speech corpus, introduced by Akbayan Bekarystankyzy et al., contains around 332 hours of transcribed audio and more than 153,000 utterances from participants representing diverse regions, age groups, and genders [60]. The KSC has undergone rigorous evaluation by native Kazakh speakers to ensure it high quality; it stands as the largest publicly available database for advancing Kazakh speech and language processing applications. The authors detail the procedures for data collection and preprocessing, describe the database specifications,

and share their experiences along with challenges encountered during its development. Initial speech recognition experiments conducted on the KSC have shown promising results with a character error rate of 2.8% and a word error rate of 8.7% on the test set. Moreover, an Asp net recipe for the speech recognition model has been released to facilitate experiment reproducibility and simplify corpus usage.

An alternative to CTC is the Sequence-to-Sequence with attention model, which includes both an encoder and a decoder [61]. The encoder compresses the audio frame data into a more efficient representation by reducing the number of neurons in the layers, while the decoder, together with a recurrent neural network, reconstructs words, symbols and phonemes based on this compressed representation.Soltau and other [6] experts trained the model to recognize phonemes depending on the context, using CTC to create subtitles for YouTube videos. In sequential models, recognition is about 13-35% faster compared to basic systems. The adaptation of CTC models is a transformative RNN that combines two RNNs into a specific system: one network works similarly to the CTC network, processing each time period processed by the input sequence, taking into account the previous ones. Dynamic programming is used in CTC networks to calculate algorithms, as well as forward and reverse transition algorithms, while eliminating the limitations present in both RNN networks.

Nurgali Kadyrbek and his colleagues conducted a study on the transcription of spoken Kazakh in rapidly changing circumstances [62]. The research is devoted to the phonetic composition of the Kazakh language, the technical aspects of creating a database of transcribed audio recordings and the use of deep neural networks to represent speech. With a high-quality collection of transcribed audio recordings totaling 554 hours, this study provides valuable information on the frequency of letters and syllables, as well as demographic variables such as gender, age and geographic location of native speakers. The extensive vocabulary included in the dataset is an important resource for the development of speech-related modules. Machine learning experiments were conducted using the DeepSpeech2 model with a sequence-by-sequence architecture, which combines an encoder and decoder control mechanism to increase the reliability of filters and reduce dependence on accurate positioning of objects on the map by implementing at the character level. In addition, during the training process, both supervised and unsupervised learning methods combined to reduce the weight of the model by 66.7% while maintaining relative accuracy.The analysis of test samples showed a decrease in the error rate when entering characters by 7.6% compared to existing models, which demonstrates advanced capabilities. The proposed design allows it to be used on platforms with limited capabilities.

Weijing Meng and colleagues [63] discuss the challenges of developing accurate speech recognition systems for low-resource languages such as Kazakh. They emphasize the potential of unsupervised pre-training to improve performance and introduce wav2vec-F, a model that utilizes unsupervised pre-training to learn speech representations from unlabeled audio data on a large scale. In addition, they incorporate a factorized TDNN layer to better capture the relationship between voice and time step before and after quantization. The authors also integrate speech synthesis to enhance speech recognition performance. Their experiments show

that wav2vec-F effectively leverages unlabeled data from non-target languages, demonstrating that multi-language pre-training outperforms single-language pre-training. By employing multi-language combined with TTS, the proposed model achieves comparable results as previous end-to-end models using only a small set of labeled Kazakh speech data.

Musakhodzhayeva et al. [64] expanded the corpus for the synthesis of Kazakh text into speech in order to overcome difficulties in developing high-quality TTS systems for the Kazakh language. The volume of the corpus was increased from 93 hours to 271 hours, including additional speakers and covering various topics. The authors consider linguistic problems characteristic of the agglutinative nature of the Kazakh language and offer detailed information on the creation, training and evaluation of the corpus. The effectiveness of this corpus is demonstrated by the example of achieving a subjective average score exceeding 3.6 for all five native speakers when building reliable TTS models. Posting this resource on GitHub makes a significant contribution to speech and language research not only for Kazakh, but also for other Turkic languages for which there are insufficient resources.

Khomitsevich and co-authors have implemented a speech recognition and synthesis system in both Kazakh and Russian, taking into account the prevalence of bilingualism among native speakers of the Kazakh language [65]. The development process included the creation of a text processing and transcription system capable of working with both languages, resulting in a dual-purpose system supporting applications for speech synthesis and recognition in both Kazakh and Russian. In addition, the authors used recordings of a single bilingual speaker to create a Kazakh text-to-speech and an additional Russian voice to ensure consistency between the languages. This work marks the development of a universal solution adapted for native speakers of the Kazakh language, which demonstrates readiness for practical application in multilingual contexts, contributing to the development of speech technologies.

The study [52] examines the difficulties faced by automatic speech recognition systems in the Kazakh language that recognize only one language, due to widespread bilingualism among native Kazakh speakers and their tendency to mix Russian. The purpose of this article is to evaluate the advantages of bilingual learning using a matrix language along with data from exclusively spoken Kazakh, as opposed to using exclusively code-switched data. Russian russians used two separate sets of information: one containing mixed Kazakh-Russian speech, and the other consisting entirely of Russian speech without mixing languages. The study used training models designed specifically for each language dataset and a combined model using both languages. The main goal was to evaluate how well models trained in code switching performed compared to models trained in a complete sentence in both languages. The results show that bilingual learning not only improves the efficiency of the model when it comes to matrix language words, but also significantly improves the accuracy of recognition of embedded words in longer phrases or sentences.For terms with code substitution, there was a significant absolute decrease in the frequency of errors in words by 14.69%, which emphasizes the effectiveness of bilingual education in improving the reliability and accuracy

of the ASR system when working with mixed language contexts.

In the context of bilingualism, which is common among native Kazakh speakers, the practice of switching to Russian in Kazakh speech creates difficulties for ASR systems designed for monolingual Kazakh. It is often difficult for these systems to accurately transcribe Russian words included in Kazakh speech. To solve this problem, several studies have explored the benefits of bilingual learning using separate sets of monolingual data for both a matrix language and an embedded language, as opposed to using code-switched data exclusively.In a recent study, Mussakhojayeva and et al. used two different datasets: one consisting of Kazakh speech with code switching, and the other containing Russian speech without code switching [66]. They used two learning methods: they trained separate models on each dataset representing a single language, and created a bilingual model combining both datasets. The main goal was to compare how well the models trained in code-switched speech performed with those trained in full pronunciation in both languages.The results of the experiment showed that bilingual learning significantly increased the accuracy of ASR models in recognizing words from both languages, including their embedded forms. It is noteworthy that recognition of embedded Russian words has been significantly improved, and the absolute error rate in words has increased by 14.69% when working with code switching scenarios. This indicates that bilingual learning can be crucial to improve the stability and accuracy of ASR systems, especially in situations characterized by frequent switching between languages such as Kazakh and Russian.

Topic	Authors	Methodologies or key focus areas
Automatic speech recognition - a brief history of the technology development	B. Juang and Lawrence Rabine	Practical applications of ASR systems in providing real-time travel information, stock quotes, weather forecasts, and other services.
Speech recognition: A model and a program for research	Hallet and Stevens	An analysis-synthesis process used to extract a series of phonemes from a dynamically changing spectrum of the input signal
Automatic speech recognition and speech variability: A review	M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens	Highlights the influence of environmental factors like background noise, as well as the necessity to tackle speech variability stemming from foreign accents and regional difference

Continued on next page

Topic	Authors	Methodologies or key focus areas
Deep and wide: Multiple layers in automatic speech recognition	Nelson Morgan	Examines the use of discriminatively trained feedback networks in speech recognition, with a focus on multi-level processing prior to word sequence decoding using hidden Markov models
Current challenges and application of speech recognition process using natural language processing: A survey	Neha Chadha, R.C. Gangwar, and Rajeev Bed	Investigates the advancement of recognition systems that aim to improve communication between humans and machines through processing text or speech.
Natural language processing techniques in text-to-speech synthesis and automatic speech recognition	Alexandre Trilla	The examination of the connection between natural language processing and speech technologies.
Self-supervised learning with random-projection quantizer for speech recognition.	Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu	Propose a technique for autonomous speech recognition that utilizes random projection quantization. Their method involves using a model to predict labels from concealed speech signals, demonstrating similar performance to advanced models on LibriSpeech while surpassing wav2vec 2.0 and w2vBERT in certain tasks. By decoupling the quantizer from the recognition model, this approach streamlines the framework and emphasizes its effectiveness, especially for streaming models and multilingual applications.
Long short-term memory recurrent neural network for automatic speech recognition	Jane Oruh, Serestina Viriri, and Adekanmi Adegun	Improved deep learning model has been presented, utilizing recurrent neural networks and long short-term memory to tackle the difficulties of processing continuous input.
Continued on next page		

Topic	Authors	Methodologies or key focus areas
NDeep speech 2: End-to-end speech recognition in english and mandarin	Dario Amodei and et al.	Explore effectiveness of integrated deep learning methods in Multilingual Automatic Speech Recognition (ASR) for English and Mandarin Chinese.
Convolutional neural networks for speech recognition	Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu	Highlights the efficacy of mixed models that combine deep neural networks with hidden Markov models in improving speech recognition compared to conventional Gaussian mixed model and HMM systems.
Signal modeling techniques in speech recognition	J.W. Picone	An in-depth analysis is presented on the signal processing methods used in modern speech recognition systems, with a focus on commonly utilized techniques.
Long short-term memory recurrent neural network for automatic speech recognition	Jane Oruh, Serestina Viriri, and Adekanmi Adegun	Improved deep learning model has been presented, utilizing recurrent neural networks and long short-term memory to tackle the difficulties of processing continuous input.
Process speech recognition system using artificial intelligence technique	Anupam Choudhary and Ravi Kshirsagar	Explore the complexities of utilizing artificial intelligence methods for speech recognition, covering a range of models such as acoustic, language, trigram, class, source, and channel models.
An advanced nlp framework for high-quality text-to-speech synthesis	Katalin Ungurean and Dragos Burileanu	Provide a summary of the Natural Language Processing phase in the Romanian Text-to-Speech system, including its incorporation into the Speech Synthesis Markup Language system, which is a well-established standard for creating TTS documents and facilitating communication between modules.
Continued on next page		

Topic	Authors	Methodologies or key focus areas
High-quality text-to-speech synthesis : an overview	Thierry Dutoit	Provides a comprehensive discussion of the progress in speech-to-text synthesis, highlighting the importance of Digital Signal Processing and Natural Language Processing elements across different uses such as telecommunications and aiding people with disabilities
Specaugment: A simple data augmentation method for automatic speech recognitio	Daniel S. Park and et al.	Introduce improvements to ASR neural networks by incorporating special additional elements, explaining their methodology with Listen, Attend and Spell networks. These enhancements have led to substantial performance gains on the LibriSpeech and Switchboard datasets, representing a marked advancement in ASR technology.
Deep speech: Scaling up end-to-end speech recognition	Awni Hannun and et al.	Deep Voice is introduced as a comprehensive deep learning approach for voice identification, surpassing conventional systems and showing exceptional performance in challenging conditions, thus showcasing its capacity to transform voice recognition technology.
Attention-based models for speech recognition	Jan Chorowsk and et al.	Improve attention-based recurrent sequence models for phoneme recognition, demonstrating enhanced robustness and tackling the difficulties presented by input sequences of varying lengths, indicating potential wider uses beyond just speech identification.

Continued on next page

Topic	Authors	Methodologies or key focus areas
Speech to text conversion for multilingual languages	Yogita H. Ghadage and Sushama D. Shelke	Developed a multilingual system for converting speech to text, utilizing MFCC feature extraction and SVM-based classification. Their approach demonstrated significant accuracy, reaching 93.625% for Marathi, which exceeds the performance of previous techniques.
Towards unsupervised speech-to-text translation	Yu-An Chung and et al.	Presented a new system for converting speech to text by utilizing monolingual data. Their findings revealed that the VecMap system outperformed MUSE in unregulated translation situations.
Weak-attention suppression for transformer based speech recognition	Yangyang Sh and et al.	Presented a technique for reducing weak attention by dynamically modifying the focus probability to minimize excessive attention on acoustic frames, resulting in decreased word error rates when compared to conventional transformer models. There was a significant 10% improvement in WER during pure testing in Libre Speech.
End-to-end multi-speaker speech recognition with transformer	Chang X and et al.	Implemented a Transformer architecture to replace the RNN-based encoder-decoder model in the E2E system.
Very large vocabulary asr for spoken russian with syntactic and morphemic analysis	Alexey Karpov, Irina Kipyatkova, and Andrey Ronzhin.	Investigated the use of DNN confidence for continuous Russian speech recognition and utilized finite state machine-based converters for improved accuracy over hidden Markov models.
Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers	Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong.	Presents a new approach for recognizing speech in multiple languages, using an SHL-MDNN design with shared hidden layers across languages but specific softmax layers for each language.
Continued on next page		

Topic	Authors	Methodologies or key focus areas
Multi-task learning u-net for single channel speech enhancement and mask-based voice activity detection	Geon Woo Lee and Hong Kook Kim	Presents a new approach for recognizing speech in multiple languages, using an SHL-MDNN design with shared hidden layers across languages but specific softmax layers for each language.
Automatic transcription of czech, russian, and slovak spontaneous speech in the MALACH project // proceedings of eurospeech	J Psutka, P Ircing, J V Psutka, J Haji˘c, W J Byrne, and J Mirovsk	Utilized deep neural networks to identify natural CzechThe research utilized deep neural networks to identify natural Czech, Slovak, and Russian speech in interviews with Holocaust survivors
A crowdsourced open-source kazakh speech corpus and initial speech recognition baseline	J Yerbolat Khassanov and et al.	Utilized deep neural networks to identify natural CzechThe research utilized deep neural networks to identify natural Czech, Slovak, and Russian speech in interviews with Holocaust survivors
End-to-end speech recognition systems for agglutinative languages	Akbayan Bekarystankyzy and Mamyrbayev Orken	presented the Kazakh Speech Corpus, a freely available dataset containing 332 hours of transcribed audio and more than 153,000 varied utterances
The development of a kazakh speech recognition model using a convolutional neural network with fixed character level filters	Nurgali Kadyrbek, Madina Mansurova, Adai Shomanov, and Makharova Gaukhar	investigated the transformation of spoken Kazakh in dynamic conditions, with a specific focus on phonetic structure analysis, technical database development, and the use of deep neural networks for speech representation
A study of speech recognition for kazakh based on unsupervised pre-training	Weijing Meng and Nurmemet Yolwas	Investigated the transformation of spoken Kazakh in dynamic conditions, with a specific focus on phonetic structure analysis, technical database development, and the use of deep neural networks for speech representation
A study of multilingual end-to-end speech recognition for kazakh, russian, and english.	Saida Mussakhojayeva, Yerbolat Khassanov, and Huseyin Atakan Varol	Investigated the use of bilingual learning techniques by analyzing distinct monolingual datasets for Kazakh and Russian.

Continued on next page

Topic	Authors	Methodologies or key focus areas
Evaluation of the efficiency of state-of-the-art speech recognition engines.	Asma Trabelsi, Sebastien Warichet, Yassine Aajaoun, and Severine Soussilane	The study tests automatic speech recognition methods for diagnosing speech apraxia in children.
A Comparative Study of Feature Extraction Techniques for Speech Recognition System.	Deshmukh, Ratnadeep and Kurzekar, Pratik and Waghmare, Dr. Vishal and Shrishrimal, Pukhraj,	Speech recognition facilitates communication between humans and machines in various fields, such as voice-activated dialing, telecommunications, device control, and automation.
Diagnostic assessment of childhood apraxia of speech using automatic speech recognition (asr) methods	John-Paul Hosom, Lawrence Shriberg, and Jordan R Green	Several research studies have validated the similarity of findings from automated and manual analysis of DQ and QMS in detecting speech apraxia in pediatric patients.

Chapter 3

Methodology

This section of the dissertation will outline a research methodology designed to examine and contrast the efficacy of speech recognition models in the Kazakh language through VOSK-based models. It will delineate the primary aims and objectives of this investigation, along with the methodologies and approaches employed to accomplish them. The primary objective of this research is to assess and contrast two speech recognition models in the Kazakh language: VOSK big and VOSK small. To achieve this aim, the study addresses various essential activities including examining the structure of the models, choosing and organizing data for training, as well as assessing the performance of trained models through the Word Error Rate measurement.

Several techniques and strategies are employed to accomplish the study's established aims and purposes. Specifically, models will be analyzed and compared based on experiments conducted using the KazakhTTS dataset [64]. Common tools and approaches in machine learning and natural language processing will be utilized for model training and assessment. Furthermore, statistical methods and comparative analysis will be applied to scrutinize the findings as well as detect possible constraints of the models. Speech recognition is a significant area of study within artificial intelligence, with diverse practical uses such as automated transcription, voice-controlled interfaces and systems, and speech recognition technology in various languages. The unique linguistic features of the Kazakh language pose additional challenges for speech recognition due to its distinctiveness from Indo-European languages. [31]

Deep learning and machine learning techniques are currently being widely utilized for speech recognition in the Kazakh language. Deep learning, a subset of machine learning methods, involves models seeking to uncover various levels of data representations through multiple layers of information processing.

The fundamental elements within the speech recognition systems in Kazakh consist of the architectural components found within deep neural networks. These encompass activations, convolutional and recurrent layers, as well as long-term and short-term memory mechanisms [67]. For instance, recurrent neural networks are extensively employed for modeling sequential data such as audio files containing speech, while convolutional neural networks effectively extract spatial features from input data. Meanwhile, transformers exhibit exceptional performance in tasks

involving sequence processing and can be adapted to handle speech data in the Kazakh language.

3.1 Dataset

The KazakhTTS initiative was carried out with the approval of Nazarbayev University's Institutional Ethics Committee. [68]

The first step in creating the dataset involved gathering written materials. Texts were obtained from news websites in chronological order to ensure a diverse range of topics and to avoid issues commonly found in web crawlers. These texts were manually reviewed to eliminate inappropriate content, such as sensitive political matters, privacy violations, and violent themes. They were then stored in DOC format for professional announcers' convenience. In total, over 2,000 articles of varying lengths were collected. Two professional announcers—a woman and a man—were selected to narrate the texts. They were tasked with recording audio files at their natural pace and style within a quiet indoor environment while adhering to pronunciation rules and comma usage guidelines. Together they covered approximately 1,250 articles; there was an overlap of around 300 articles between them.

Five individuals who are fluent in the Kazakh language were employed to create written transcriptions and synchronize them with audio recordings. They utilized the Praat toolkit to segment the audio files into individual sentences before aligning them with the corresponding text.

The KazakhTTS collection includes approximately 93 hours of audio content, with over 42,000 individual segments. Table 2 displays the overall statistics for the dataset, while Figure 1 illustrates the distribution of segment lengths and their respective durations. Creating this dataset required approximately five months, resulting in an uncompressed dataset size of around 15 GB.

The KazakhTTS dataset is organized with the assets of two expert speakers kept in distinct directories. Each directory contains a metadata file and two subdirectories holding audio recordings and their associated transcriptions. The names of the files for both the audio recordings and transcriptions follow a consistent format, using the original article ID followed by the segment ID. However, audio files are saved in WAV format, while transcriptions are stored as TXT files using UTF-8 encoding.

The Table 3.1. presents detailed information about the content and characteristics of a specific portion of the KazakhTTS dataset, including the number of audio segments, their duration, speaker usage, and various other properties.

The Figure 3.1 illustrates the distribution of audio segment durations within the dataset. It indicates how frequently different durations occur, providing insight into the predominant duration values. For instance, a peak at 5 seconds suggests that a significant proportion of audio segments in this subset have a duration close to 10 seconds. Therefore, this graph aids in comprehending the overall distribution of audio segment durations within the analyzed dataset.

Table 3.1 – Dataset specification

Category	Values	
	Duration	Words
Total	60554 sec	3702650
Mean	8.3 sec	13
Max	35 sec	75
Min	1 sec	1

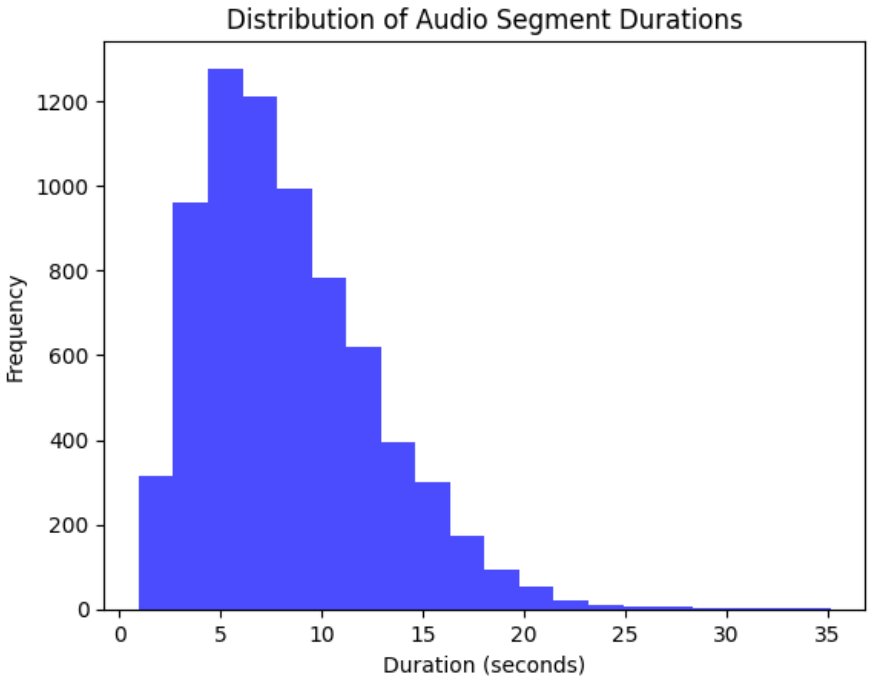

Figure 3.1 – Distribution of audio segment duration

3.2 Model Description

In my research, I was able to effectively convert audio data into text format using the powerful tool VOSK. Based on cutting-edge deep learning technology, VOSK is a powerful open-source speech recognition tool that supports 17 languages and dialects for use as models at the time of writing.VOSK provides speech recognition for chatbots, smart home devices and virtual assistants.The model was developed by Alpha Cephei Mc. This library was created in 2019. Vosk is based on Kaldi speech recognition technology, but Vosk provides a simple interface for using this technology in various applications [69]. One of the reasons VOSK was chosen was because of the uncomplicated documen- tation. Another reason in favor of VOSK was that there is an explicit Kazakh model proposed to adapt the model. It is also capable of producing film subtitles and transcripts of lectures and interviews.VOSK's versatility as a research tool is largely due to its ability to handle a wide range of languages and dialects. In addition, its real-time functionality has allowed me to process large amounts of audio data quickly and efficiently, with immediate results. Vosk is one of the open source STT engines. Vosk must be

installed before it can be used. There are two types of the Kazakh VOSK model: big and small. For comparison, I used both options in my experiment. It can be installed via pip3. The VOSK model was downloaded and initialised prior to the start of the research, along with the installation of the required libraries, including VOSK, pydub and scikit-learn. The model was set up by selecting the version based on the requirements and objectives of the research. Initializing the KaldiRecognizer recognizer with a model and sample rate indication was part of the processing process. Automatic speech recognition holds a prominent place in the fields of natural language processing and artificial intelligence. The Kaldi library is widely recognized for its comprehensive toolset, playing a pivotal role in training speech recognition models and their application across diverse uses.

The Kaldi project, created by the University of California, San Diego, was designed to offer a versatile and open platform for advancements in speech recognition research and development. It offers resources for instruction in acoustic, linguistic and phonetic modeling as well as for the processing and examination of audio data.

The developers of the Vosk library have suggested the KaldiRecognizer interface for utilizing trained models and algorithms from Kaldi to perform real-time speech recognition. This tool is highly efficient, capable of both online and offline processing of streaming audio data as well as audio files.

Key features of KaldiRecognizer include:

- KaldiRecognizer leverages the infrastructure and algorithms of Kaldi to deliver precise speech recognition.
- KaldiRecognizer has the capability to operate with diverse languages and dialects, adjusting to the specific linguistic traits of various language groups due to Kaldi's multilingual support.
- KaldiRecognizer demonstrates exceptional speed and efficiency, making it suitable for high-performance applications in real-world scenarios.
- KaldiRecognizer's flexibility allows for seamless integration into a wide range of applications and platforms, such as server and mobile systems.

KaldiRecognizer offers new prospects for developers and researchers involved in speech recognition [67]. With its effectiveness, precision, and adaptability, it has become an essential instrument across a range of disciplines including voice control systems, automated translation, and speech analysis.

After that, the audio files were read in binary format, and the recognizer received the audio data from them. The identification outcome was stored for a future comparison with the real transcription that was taken out of the relevant files. Finding, sending, and analyzing audio data were among the processing procedures. The outcomes were then compared to accurate transcriptions. Each audio file underwent this procedure in order to assess the accuracy of recognition. Therefore, analyzing the performance of a speech recognition model using the VOSK library provides a useful perspective on the use of automated speech recognition technologies. The actual and predicted transcriptions were analyzed using the WER method to calculate the degree of consistency between them. The WER results enable us to assess the effectiveness of the Vosk model across different languages and regional speech variations. Low WER scores suggest strong

accuracy in speech recognition, whereas higher scores may signal a requirement for additional adjustments or enhancements to the model. The Word Error Rate [70] serves as a metric to assess the performance of speech recognition systems. It quantifies the level of disparity between the true transcription (ground truth) and the output from the speech recognition system. WER is typically presented as a percentage, indicating the proportion of words in phonetic transcription that were inaccurately identified or missed entirely. WER can be formally defined as the ratio of the combined number of word insertions, deletions, and substitutions to the total word count in the original transcription:

$$\text{WER} = \frac{\text{Ins} + \text{Del} + \text{Sub}}{\text{N}} \times 100\% \tag{3.2.1}$$

Where:

- Ins is the number of words that were incorrectly inserted (not present in the actual transcription, but present in the predicted one)
- Del is the number of words that were incorrectly deleted (present in the actual transcription, but absent in the predicted one)
- Sub is the number of words that were incorrectly replaced (present in both transcriptions, but different among themselves)
- N is the total number of words in the actual transcription.

The lower the WER value, the better the quality of the speech recognition system. For example, a WER of 0% means that the predicted transcription is identical to the actual one, and a WER of 100% means that all words in the predicted transcription differ from the actual one.

WER is an important metric in evaluating the quality of speech recognition systems and is used in various fields, including research, commercial application development, and performance evaluation of existing systems. The results can be used as a basis for further study and development in this area, providing new insights into audio processing.

Chapter 4

Results

4.1 Settings for experimentation

1. Opening an audio file: The audio file was opened for reading in binary format. Then the sampling rate (sampling rate) was obtained from the audio file.
2. Initializing the speech recognizer: The KaldiRecognizer recognizer object was created using the specified model and the resulting audio file sampling rate.
3. An important aspect when initializing the recognizer was also to take into account the number of channels of the audio file. To do this, you could explicitly specify the number of channels when initializing the recognizer.

4.2 Word Error Rate

The primary goal of this research was to assess and contrast the performance of two voice recognition models—VOSK big and VOSK small—for the Kazakh language. On a dataset made up of Kazakh speech samples from the KazakhTTS dataset, the efficacy of these models was assessed using the word error rate index (WER).

The results of our experiments revealed the following WER scores for the two models:

VOSK small: This model showed a slightly better performance with a WER of 53%. This indicates that 53% of the words in the transcribed text were either substituted, inserted, or deleted incorrectly when compared to the reference text.

VOSK big: The model achieved a WER of 51%. This suggests that the smaller model made fewer errors in transcription, resulting in a slightly higher accuracy compared to the VOSK small model.

A comparison of the WER indicators between the two models shows that VOSK big is slightly superior to VOSK small in terms of accuracy for the Kazakh language. This difference in performance can be influenced by several factors, including the architecture of the models and how they handle the specific phonetic and grammatical features of the Kazakh language.

4.3 Code execution time

During the study, it was found that the completion time of the speech recognition process using the VOSK small model is approximately 17-18 minutes, while for the VOSK big model this time is approximately 20-21 minutes. These results indicate differences in performance between the two speech recognition models in the context of audio file processing.

The described time indicators reflect the time required to complete the speech recognition process, including preprocessing of audio data, performing computational operations and generating text output. According to the results of the study, it turned out that the VOSK small model demonstrates a faster completion rate compared to the VOSK big model.

4.4 Hardware resources

The following are the main hardware characteristics that influenced the execution and results of the study:

- Graphics Processing Unit (GPU): The integrated Intel(R) UHD Graphics provided by the Intel Core(TM) i3-1005G1 processor was used. This GPU provides basic graphics processing on a laptop.
- Processor (CPU): The study used an Intel(R) Core(TM) i3-1005G1 processor with a clock frequency of 1.20 GHz. This processor is dual-core and provides sufficient computing power to execute algorithms and software related to speech processing.
- RAM: The amount of RAM is 128 GB. This amount of RAM provides the necessary amount of resources to perform research tasks, such as processing and analyzing audio data and running machine learning algorithms.
- specified hardware characteristics of the Acer Aspire 5 A515-55-55V2 laptop provided the necessary computing power and resources to perform research and achieve the goals set.

Chapter 5

Discussion

The results of our experiments showed the following WER values for the two models. The VOSK small model, which is 42Mb in size, demonstrated a WER of 53%, indicating that 53% of the words in the transcribed text were either replaced, inserted, or deleted incorrectly compared to the reference text. The VOSK big model, which is 378Mb in size, achieved a WER of 51%, indicating that the large model made fewer transcription errors, providing slightly higher accuracy compared to the VOSK small model.

Despite the smaller size and presumably less intensive preparation of the VOSK small model, its effectiveness turned out to be approximately the same as that of the VOSK big model. Also, during the study, it was found that the completion time of the speech recognition process using the VOSK small model is approximately 17-18 minutes, while for the VOSK big model this time is about 20-21 minutes. These results indicate performance differences between the two speech recognition models in the context of audio file processing. The results of the experiment showed that both the VOSK big and VOSK small models face certain limitations in their ability to accurately recognize Kazakh speech. One of the key disadvantages of both models is the difficulty in correctly identifying the endings of words. This leads to incomplete or incorrect representations of the final sounds in some cases, which can significantly affect the understanding and accuracy of the transcribed text. Moreover, both models showed cases of incorrect word recognition. These errors may be related to various factors, such as variations in pronunciation, dialects, accents, or other linguistic features that were not fully taken into account in the learning process of the model. As a result of such inaccuracies, the accuracy and reliability of the transcribed output decreases, which complicates the use of models for real-world applications where speech fidelity plays an important role.

These limitations indicate the need for further improvement of the VOSK big and VOSK small models, including by improving their adaptation to the specific features of the Kazakh language and refining speech recognition algorithms. It is also important to carefully select and prepare training data so that the models are more resistant to various variations in user speech.

Chapter 6

Conclusions and future work

6.1 Conclusions

Ultimately, our research findings have verified that there are substantial constraints in accurately recognizing Kazakh speech in both the VOSK big and VOSK small models. While these models exhibit variations in performance and size, they encounter similar challenges, including struggles in identifying word endings and instances of incorrect word recognition.

This highlights the necessity for enhancing speech recognition techniques and models to attain greater precision and dependability in the realm of Kazakh language. It is crucial to persist in investigating this field to create more efficient and linguistically tailored speech recognition models.

In addition, the results also highlight the importance of diverse and high-quality training data for model training, as well as the need for systematic analysis and improvement of speech processing and recognition algorithms.

Overall, this study provides valuable insights and indicates the need for further efforts in the development and improvement of speech recognition systems for more effective use in real-world applications and situations.

6.2 Future work

The experiment brought to light certain limitations in the VOSK big and VOSK small models' ability to accurately transcribe Kazakh speech. Specifically, challenges were observed in both models' capability to correctly identify word endings and in instances where words were inaccurately recognized. These limitations could potentially impact the overall comprehension and accuracy of the transcribed text, which is crucial for effective communication and understanding.

To address the limitations of the VOSK big and VOSK small models in Kazakh speech recognition, several potential directions for future research and enhancements have been suggested. One promising approach involves further refining the VOSK models using a broader and more varied collection of audio recordings of the Kazakh language. Incorporating a wider range of linguistic variations, accents, and speech patterns in training data has the potential to enhance the models' capability to recognize word endings and accommodate pronunciation differences more

efficiently.

Future research could explore the integration of phonetic variations into the learning process of models. Considering phonetic differences and linguistic subtleties during the learning stage can enhance models' understanding of speech patterns in the Kazakh language, resulting in more precise transcriptions.

A thorough examination of the output data from VOSK models can yield valuable insights into instances of inaccurate recognition and errors. Using this information, potential corrective actions may be taken in the future, including refining phonetic patterns, modifying recognition thresholds, or creating post-processing algorithms to rectify recognized text inaccuracies.

In upcoming research, adhering to these guidelines will enable scientists to overcome the recognized shortcomings of VOSK models and progress Kazakh speech recognition technologies. This advancement will ultimately enhance the accessibility and utility of automatic speech recognition systems for native Kazakh speakers.

Bibliography

[1] Marianne Engen Matre and David Lansing Cameron. A scoping review on the use of speech-to-text technology for adolescents with learning difficulties in secondary education. *Disability and Rehabilitation: Assistive Technology*, 19(3):1103–1116, 2024. doi: 10.1080/17483107.2022.2149865.

[2] Maria Kambouri, Helen Simon, and Greg Brooks. Using speech-to-text technology to empower young writers with special educational needs. *Research in Developmental Disabilities*, 135:104466, 2023. ISSN 0891-4222. doi: https://doi.org/10.1016/j.ridd.2023.104466. URL https://www.sciencedirect.com/science/article/pii/S0891422223000446.

[3] Nelson Morgan. Deep and wide: Multiple layers in automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1): 7–13, 2012. doi: 10.1109/TASL.2011.2116010.

[4] Thierry Dutoit. High-quality text-to-speech synthesis : an overview. 2004. URL https://api.semanticscholar.org/CorpusID:10693488.

[5] Antonio Andriella, Raquel Ros, Yoav Ellinson, Sharon Gannot, and Séverin Lemaignan. Dataset and evaluation of automatic speech recognition for multilingual intent recognition on social robots. page 865–869, 2024. doi: 10.1145/3610977.3637473. URL https://doi.org/10.1145/3610977.3637473.

[6] C Englund. Speech recognition in the JAS 39 gripen aircraft-adaptation to speech at different g-loads. *Technology*, 2004.

[7] Maghilnan S and Rajesh M. Sentiment analysis on speaker specific speech data. 02 2018.

[8] Ryan Whetten and Casey Kennington. Evaluating and improving automatic speech recognition using severity. pages 79–91, July 2023. doi: 10.18653/v1/2023.bionlp-1.6. URL https://aclanthology.org/2023.bionlp-1.6.

[9] Rafael Luque, Adrián R Galisteo, Paloma Vega, and Eduardo Ferrera. SIMO: An automatic speech recognition system for paperless manufactures. In *10th Manufacturing Engineering Society International Conference (MESIC 2023)*, Switzerland, October 2023. Trans Tech Publications Ltd.

[10] Ratnadeep Deshmukh, Pratik Kurzekar, Dr. Vishal Waghmare, and Pukhraj Shrishrimal. A comparative study of feature extraction techniques for speech recognition system. *International Journal of Innovative Research in Sci-*

ence, Engineering and Technology, 3:18006–18016, 12 2014. doi: 10.15680/IJIRSET.2014.0312034.

[11] John-Paul Hosom, Lawrence Shriberg, and Jordan Green. Diagnostic assessment of childhood apraxia of speech using automatic speech recognition (asr) methods. *Journal of medical speech-language pathology*, 12:167–171, 01 2005.

[12] Jianliang Meng, Junwei Zhang, and Haoquan Zhao. Overview of the speech recognition technology. pages 199–202, 2012. doi: 10.1109/ICCIS.2012.202.

[13] Zixing Zhang, Jürgen Geiger, Jouni Pohjalainen, Amr El-Desoky Mousa, Wenyu Jin, and Björn Schuller. Deep learning for environmentally robust speech recognition: An overview of recent developments. 2018.

[14] Naomi Caselli and Ariel Cohen-Goldberg. Lexical access in sign language: a computational model. *Frontiers in Psychology*, 5:428, 05 2014. doi: 10.3389/fpsyg.2014.00428.

[15] Denis Arnold, Fabian Tomaschek, Konstantin Sering, Florence Lopez, and R Harald Baayen. Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLoS One*, 12(4):e0174623, April 2017.

[16] J. L. Morgan. *Two Types of Convention in Indirect Speech Acts*, pages 261 – 280. Brill, Leiden, The Netherlands, 1978. ISBN 9789004368873. doi: 10.1163/9789004368873_010. URL https://brill.com/view/book/edcoll/9789004368873/BP000010.xml.

[17] Guus de Krom. Consistency and reliability of voice quality ratings for different types of speech fragments. *J. Speech Lang. Hear. Res.*, 37(5):985–1000, October 1994.

[18] Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, 128:32–37, 2018. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2018.03.005. URL https://www.sciencedirect.com/science/article/pii/S1877050918302187. 1st International Conference on Natural Language and Speech Processing.

[19] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, and C. Cox. Asr for emotional speech: Clarifying the issues and enhancing performance. *Neural Networks*, 18(4):437–444, 2005. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2005.03.008. URL https://www.sciencedirect.com/science/article/pii/S0893608005000419. Emotion and Brain.

[20] Naoki Hirayama, Koichiro Yoshino, Katsutoshi Itoyama, Shinsuke Mori, and Hiroshi G. Okuno. Automatic speech recognition for mixed dialect utterances by mixing dialect language models. *IEEE/ACM Transactions*

on Audio, Speech, and Language Processing, 23(2):373–382, 2015. doi: 10.1109/TASLP.2014.2387414.

[21] Larry E Humes and Lisa Roberts. Speech-recognition difficulties of the hearing-impaired elderly. *J. Speech Lang. Hear. Res.*, 33(4):726–735, December 1990.

[22] Dong Wang, Xiaodong Wang, and Shaohe Lv. An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8), 2019. ISSN 2073-8994. doi: 10.3390/sym11081018. URL https://www.mdpi.com/2073-8994/11/8/1018.

[23] Kenneth N Stevens. Toward a model for speech recognition. *J. Acoust. Soc. Am.*, 32(1):47–55, January 1960.

[24] Martín Haro, Joan Serrà, Álvaro Corral, and Perfecto Herrera. Power-law distribution in encoded mfcc frames of speech, music, and environmental sound signals. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, page 895–902, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312301. doi: 10.1145/2187980.2188220. URL https://doi.org/10.1145/2187980.2188220.

[25] B. H. Juang and L. R. Rabiner. Hidden markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991. doi: 10.1080/00401706.1991.10484833.

[26] Lux Li, Robert Rehr, Patrick Bruns, Timo Gerkmann, and Brigitte Röder. A survey on probabilistic models in human perception and machines. *Front. Robot. AI*, 7:85, July 2020.

[27] Jen-Tzung Chien. Chapter 7 - deep neural network. In Jen-Tzung Chien, editor, *Source Separation and Machine Learning*, pages 259–320. Academic Press, 2019. ISBN 978-0-12-817796-9. doi: https://doi.org/10.1016/B978-0-12-804566-4.00019-X. URL https://www.sciencedirect.com/science/article/pii/B978012804566400019X.

[28] Imran Sheikh, Emmanuel Vincent, and Irina Illina. Training rnn language models on uncertain asr hypotheses in limited data scenarios. *Computer Speech Language*, 83:101555, 2024. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2023.101555. URL https://www.sciencedirect.com/science/article/pii/S0885230823000748.

[29] B. Juang and Lawrence Rabiner. Automatic speech recognition - a brief history of the technology development. 01 2005.

[30] M. Halle and K. Stevens. Speech recognition: A model and a program for research. *IRE Transactions on Information Theory*, 8(2):155–159, 1962. doi: 10.1109/TIT.1962.1057686.

[31] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. Automatic speech recognition and speech variability: A review. *IRE Transactions on Information Theory*, 49(10):763–786, 2007. doi: https://doi.org/10.1016/j.specom.2007.02.006.

[32] Neha Chadha, R.C. Gangwar, and Rajeev Bedi. Current challenges and application of speech recognition process using natural language processing: A survey. *International Journal of Computer Applications*, 131:28–31, 12 2015. doi: 10.5120/ijca2015907471.

[33] Alexandre Trilla. Natural language processing techniques in text-to-speech synthesis and automatic speech recognition. 2009. URL https://api.semanticscholar.org/CorpusID:17312105.

[34] Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. Self-supervised learning with random-projection quantizer for speech recognition. 2022.

[35] Jane Oruh, Serestina Viriri, and Adekanmi Adegun. Long short-term memory recurrent neural network for automatic speech recognition. *IEEE Access*, 10: 30069–30079, 2022. doi: 10.1109/ACCESS.2022.3159339.

[36] Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, and Zhenyao Zhu. Deep speech 2: End-to-end speech recognition in english and mandarin. 2015.

[37] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10): 1533–1545, 2014. doi: 10.1109/TASLP.2014.2339736.

[38] J.W. Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1215–1247, 1993. doi: 10.1109/5.237532.

[39] Anupam Choudhary. Process speech recognition system using artificial intelligence technique. 2012. URL https://api.semanticscholar.org/CorpusID:212607769.

[40] Catalin Ungurean and Dragos Burileanu. An advanced nlp framework for high-quality text-to-speech synthesis. pages 1–6, 2011. doi: 10.1109/SPED.2011.5940733.

[41] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. September 2019. doi: 10.21437/interspeech.2019-2680. URL http://dx.doi.org/10.21437/Interspeech.2019-2680.

[42] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam

Coates, and Andrew Y. Ng. Deep speech: Scaling up end-to-end speech recognition. 2014.

[43] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. 2015.

[44] Yogita H. Ghadage and Sushama D. Shelke. Speech to text conversion for multilingual languages. pages 0236–0240, 2016. doi: 10.1109/ICCSP.2016.7754130.

[45] Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. Towards unsupervised speech-to-text translation. 2018.

[46] Shigeki Karita, Nelson Enrique Yalta Soplin, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. September 2019.

[47] Yangyang Shi, Yongqiang Wang, Chunyang Wu, Christian Fuegen, Frank Zhang, Duc Le, Ching-Feng Yeh, and Michael L. Seltzer. Weak-attention suppression for transformer based speech recognition. 2020.

[48] Xuankai Chang, Wangyou Zhang, Yanmin Qian, Jonathan Le Roux, and Shinji Watanabe. End-to-end multi-speaker speech recognition with transformer. May 2020.

[49] Alexey Karpov, Irina Kipyatkova, and Andrey Ronzhin. Very large vocabulary asr for spoken russian with syntactic and morphemic analysis. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 3161–3164, 08 2011. doi: 10.21437/Interspeech.2011-791.

[50] Romain Serizel and Diego Giuliani. Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. December 2014.

[51] Hongbing Hu and Stephen Zahorian. Dimensionality reduction methods for hmm phonetic recognition. pages 4854–4857, 01 2010. doi: 10.1109/ICASSP.2010.5495130.

[52] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. pages 7304–7308, 2013. doi: 10.1109/ICASSP.2013.6639081.

[53] Olzhas Makhambetov, Aibek Makazhanov, Zhandos Yessenbayev, Bakhyt Matkarimov, Islam Sabyrgaliyev, and Anuar Sharafudinov. Assembling the kazakh language corpus. pages 1022–1031, 2013.

[54] Orken Mamyrbayev, Keylan Alimhan, Bagashar Zhumazhanov, Tolganay Turdalykyzy, and Farida Gusmanova. End-to-end speech recognition in agglutinative languages. pages 391–401, 2020.

[55] Olga Khomitsevich, Valentin Mendelev, Natalia A Tomashenko, Sergey Rybin, Ivan Medennikov, and Saule Kudubayeva. A bilingual kazakhrussian system for automatic speech recognition and synthesis. In *Proc. of the 17th International Conference on Speech and Computer (SPECOM)*, volume 9319, pages 25–33. Athens, Greece, 2015.

[56] Ying Shi, Askar Hamdullah, Zhiyuan Tang, Dong Wang, and Thomas Fang Zheng. A free kazakh speech database and a speech recognition baseline. December 2017.

[57] Geon Woo Lee and Hong Kook Kim. Multi-task learning u-net for single-channel speech enhancement and mask-based voice activity detection. *Applied Sciences*, 10(9), 2020. ISSN 2076-3417. doi: 10.3390/app10093230. URL https://www.mdpi.com/2076-3417/10/9/3230.

[58] J Psutka, P Ircing, J V Psutka, J Hajič, W J Byrne, and J Mirovsky. Automatic transcription of czech, russian, and slovak spontaneous speech in the MALACH project // proceedings of eurospeech. *Lisboa. Portugal*, pages 1349–1352, 2005.

[59] Yerbolat Khassanov, Saida Mussakhojayeva, Almas Mirzakhmetov, Alen Adiyev, Mukhamet Nurpeiissov, and Atakan Varol. A crowdsourced open-source kazakh speech corpus and initial speech recognition baseline. pages 697–706, 01 2021. doi: 10.18653/v1/2021.eacl-main.58.

[60] Akbayan Bekarystankyzy and Mamyrbayev Orken. nd-to-end speech recognition systems for agglutinative languages. *Scientific Journal of Astana IT University*, pages 86–92, 03 2023. doi: 10.37943/13IMII7575.

[61] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.*, 29(6):82–97, November 2012.

[62] Nurgali Kadyrbek, Madina Mansurova, Adai Shomanov, and Makharova Gaukhar. The development of a kazakh speech recognition model using a convolutional neural network with fixed character level filters. *Big Data and Cognitive Computing*, 7:132, 07 2023. doi: 10.3390/bdcc7030132.

[63] Weijing Meng and Nurmemet Yolwas. A study of speech recognition for kazakh based on unsupervised pre-training. *Sensors (Basel)*, 23(2), January 2023.

[64] S Mussakhojayeva, Y Khassanov, and H A Varol. Extending the Open-Source kazakh TTS corpus with more Data,Speakers, and topics. pages 5404–5411, 2022.

[65] Olga Khomitsevich, Valentin Mendelev, Natalia Tomashenko, Sergey Rybin, Ivan Medennikov, and Saule Kudubayeva. A bilingual kazakh-russian system for automatic speech recognition and synthesis. pages 25–33, 2015.

[66] Saida Mussakhojayeva, Yerbolat Khassanov, and Huseyin Atakan Varol. A study of multilingual end-to-end speech recognition for kazakh, russian, and english. page 448–459, 2021. doi: 10.1007/978-3-030-87802-3_41. URL https://doi.org/10.1007/978-3-030-87802-3_41.

[67] Asma Trabelsi, Sébastien Warichet, Yassine Aajaoun, and Séverine Soussilane. Evaluation of the efficiency of state-of-the-art speech recognition engines. *Procedia Computer Science*, 207:2242–2252, 2022. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2022.09.534. URL https://www.sciencedirect.com/science/article/pii/S1877050922014338. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 26th International Conference KES2022.

[68] Saida Mussakhojayeva, Aigerim Janaliyeva, Almas Mirzakhmetov, Yerbolat Khassanov, and Huseyin Atakan Varol. Kazakhtts: An open-source kazakh text-to-speech synthesis dataset. In *Interspeech 2021*. ISCA, August 2021. doi: 10.21437/interspeech.2021-2124. URL http://dx.doi.org/10.21437/Interspeech.2021-2124.

[69] Alpha Cephei. Vosk offline speech recognition API, Year of publication, e.g., 2024. URL https://alphacephei.com/vosk/.

[70] E. Balestrieri, M. Catelani, L. Ciani, S. Rapuano, and A. Zanobini. Word error rate measurement uncertainty estimation in digitizing waveform recorders. *Measurement*, 46(1):572–581, 2013. ISSN 0263-2241. doi: https://doi.org/10.1016/j.measurement.2012.08.016. URL https://www.sciencedirect.com/science/article/pii/S0263224112003119.