Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

DOI: 10.37943/22SNOK5872

Arypzhan Aben
Master of Science, Department of Computer Engineering,
Arypzhan.aben@ayu.edu.kz, orcid.org/0000-0001-8534-3288
Khoja Akhmet Yassawi International Kazakh-Turkish University,
Kazakhstan

Gulnur Kazbekova*
Candidate of Technical Sciences, Associate Professor, Head of the Department of Computer Engineering
gulnur.kazbekova@ayu.edu.kz, orcid.org/0000-0002-2756-7926
Khoja Akhmet Yassawi International Kazakh-Turkish University,
Kazakhstan

Zhuldyz Ismagulova
Candidate of Technical Sciences, Associate Professor,
Department of Information and Communication Technologies
zhu.ismagulova@alt.edu.kz, orcid.org/0000-0002-0979-0243
ALT University, Kazakhstan

Gulmira Ibrayeva
Candidate of Physical and Mathematical Sciences, Head of the Department
gulmira_ibraeva@mail.ru, orcid.org/0000-0001-9228-0924
Military Institute of the Air Defense Forces named after Twice Hero of the Soviet Union T.Ya. Bigeldinov

AUDIO-TO-TEXT TRANSLATION FOR THE HARD OF HEARING: A WHISPER MODEL-BASED STUDY

Abstract: This study investigates the effectiveness of the Whisper model for audio-to-text transcription, specifically targeting the enhancement of accessibility for individuals with hearing impairments. The research focuses on the processing of audio recordings obtained from WhatsApp messenger, which often contain significant background noise that complicates speech recognition. To address this issue, advanced audio processing techniques were employed, including the use of the Librosa library and the Noisereduce package for noise reduction. The spectral gating methods applied in this study effectively diminished wind noise and other ambient sounds, allowing for clearer recognition of spoken content. To ensure the quality of the processed audio, we assessed its clarity using a SimpleRNN model. The training results demonstrated a progressive reduction in loss values across epochs, confirming the successful enhancement of audio quality. Once the audio files were adequately cleaned, we utilized the Whisper model, a sophisticated machine learning tool for speech recognition developed by OpenAI, to transcribe the audio into text. The transcription process yielded accurate Kazakh language output, despite the initial challenges posed by background noise. These findings underscore the critical role of high-quality audio input in achieving reliable transcription results and highlight the potential of machine learning technologies in improving communication access for hearing-impaired individuals. This study concludes with recommendations for future research, including the exploration of additional noise reduction techniques and the application of the Whisper model across various languages and dialects. Such advancements

Received: 15.11.2024 Accepted: 25.06.2025 Published: 30.06.2025

DOI: 10.37943/22SNOK5872

could significantly contribute to creating more inclusive digital environments and enhancing the overall user experience for individuals with hearing impairments.

Keywords: Whisper model, Audio-to-text transcription, Hearing impairments, Machine learning.

Introduction

Effective communication is the foundation of human interaction; however, for millions of people worldwide, engaging in verbal communication poses significant challenges due to hearing impairments. According to the World Health Organization, by 2050, approximately 2.5 billion people will experience some degree of hearing loss, with at least 700 million requiring rehabilitation services [1]. This alarming statistic underscores the urgent need for solutions that facilitate communication for those with hearing impairments, particularly as society becomes increasingly reliant on auditory information exchange [2].

In the modern world, instant messaging platforms like WhatsApp, Telegram, and Viber have revolutionized communication by enabling rapid and continuous information exchange. These platforms serve as the basis for personal and professional interactions, allowing users to easily share text, images, and audio messages [3]. Among these features, audio messaging has gained popularity due to its convenience, particularly in situations where conveying complex information verbally is challenging or typing is impractical [4]. However, this shift to voice-based communication presents accessibility challenges for individuals with hearing impairments, who struggle to comprehend audio messages without assistance [5].

Despite advancements in communication technologies, individuals with hearing impairments are often marginalized from fully participating in these digital conversations [6]. While traditional assistive technologies, such as hearing aids or cochlear implants, can be beneficial, they do not always provide a comprehensive solution, especially in noisy environments or in cases of profound hearing loss. Furthermore, these devices are often expensive and may not be accessible to all who need them, particularly in developing regions [7]. Thus, the growing dependence on audio-based communication necessitates alternative approaches to prevent leaving behind the hearing-impaired community.

Recent advancements in artificial intelligence (AI) and natural language processing (NLP) have opened up new avenues for improving accessibility. One of the most promising solutions is the development of automatic speech recognition (ASR) models that can convert spoken language into written text in real-time. Such technology enables individuals with hearing impairments to read transcripts of audio messages or live conversations, thereby enhancing their ability to participate in communication.

This study investigates the application of the Whisper ASR model, developed by OpenAI, in transcribing audio messages in the Kazakh language. The strengths of Whisper lie in its robustness in processing background noise, accents, and variations in speech dynamics, making it an excellent candidate for transcribing real-world conversations in various languages. The model's ability to support multiple languages and dialects is particularly relevant for underrepresented languages like Kazakh, which is spoken by over 10 million people but often overlooked in global speech-to-text solutions.

The primary aim of this study is to evaluate the effectiveness of Whisper in transcribing Kazakh audio messages from instant messaging platforms [8]. It focuses on platforms widely used for daily communication in Kazakhstan, such as WhatsApp, Telegram, and Viber. By converting spoken Kazakh into text, Whisper has the potential to provide critical services for individuals with hearing impairments, granting them access to audio content that would otherwise be unavailable.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

This research not only assesses the accuracy and usability of Whisper in the context of Kazakh transcription but also highlights the broader impact of AI-based solutions in addressing the growing global issue of hearing loss. As the number of individuals with hearing impairments continues to rise, particularly with the projected increase by 2050, technologies like Whisper could play a crucial role in fostering inclusive communication environments. By enhancing accessibility in daily digital communications, this study aims to contribute to the development of assistive technologies that help bridge the gap for individuals with hearing impairments, ensuring their connection in a sound-dependent world [9].

Literature review

In the field of speech-to-text translation, recent advancements have focused on enhancing accuracy, speed, and cross-lingual capabilities, particularly to assist those with hearing impairments. Various algorithms and models have been explored to improve both automatic speech recognition (ASR) and speech-to-text systems. A wide range of research efforts have aimed to bridge the gap between different modalities such as speech and text, ensuring that these models are robust, efficient, and adaptable to diverse use cases.

One significant contribution is by Wang et al. [10], who introduced Fairseq S2T, a fast and efficient speech-to-text model using the Fairseq toolkit. Xu et al. provided a comprehensive overview of recent advances in direct speech-to-text translation, focusing on multilingualism and cross-domain robustness [11]. Other studies, such as those by Guo et al., investigated controllable text-to-speech using PromptTTS, a model integrating speech descriptions into text-based tasks [12].

The following Table 1 provides an overview of the key contributions in the field, outlining the author(s), publication year, article title, the model or algorithm proposed, and the primary research area.

Table 1. Research in the field of speech to text

Author(s) and Year	Article Title	Model or Algorithm	Research Area	Metrics	Values
Wang et al. (2020)[10]	Fairseq S2T: Fast speech-to-text modeling with Fairseq	Fairseq S2T	Speech-to-Text Modeling	WER (Word Error Rate)	12.8%
Xu et al. (2023)[11]	Recent advances in direct speech-to-text translation	N/A	Speech-to-Text Translation	BLEU	26.5
Guo et al. (2023)[12]	Promptttts: Controllable text-to-speech with text descriptions	PromptTTS	Text-to-Speech, Speech Descriptions	MOS (Mean Opinion Score)	4.37
Bapna et al. (2022)[13]	mslam: Massively multilingual joint pre-training for speech and text	mSLAM	Multilingual Speech and Text Pre-training	WER, BLEU	15.3%, 28.4
Liu et al. (2020)[14]	Bridging the modality gap for speech-to-text translation	N/A	Speech-to-Text Translation	BLEU	25.8
Tang et al. (2021)[15]	A general multi-task learning framework to leverage text data for speech-to-text tasks	Multi-Task Learning Framework	Speech-to-Text Translation, Multi-task	BLEU	27.3
Matre & Cameron (2024)[16]	A scoping review on the use of speech-to-text technology for adolescents with learning difficulties	N/A	Assistive Technology for Learning	N/A	N/A
Wu et al. (2023)[17]	On decoder-only architecture for speech-to-text and large language model integration	Decoder-Only Architecture	Speech-to-Text, Language Model Integration	WER	10.5%

DOI: 10.37943/22SNOK5872

Author(s) and Year	Article Title	Model or Algorithm	Research Area	Metrics	Values
Higuchi et al. (2021) [18]	A comparative study on non-autoregressive modelings for speech-to-text generation	Non-Autoregressive Modeling	Speech-to-Text Generation	BLEU	24.9
Wang et al. (2023)[19]	Slm: Bridge the thin gap between speech and text foundation models	SLM	Speech and Text Foundation Models	BLEU, WER	30.2, 9.4%
Bhandari et al. (2023) [20]	Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from Russia-Ukraine conflict	N/A	Multimodal Hate Speech Analysis	Precision, Recall	81.5%, 79.8%
Bano et al. (2020)[21]	Speech-to-text translation enabling multilingualism	N/A	Speech-to-Text, Multilingualism	BLEU	23.7

These studies underscore the diversity of approaches to speech-to-text modeling, from leveraging multilingual capabilities to integrating deep learning frameworks, and exploring novel architectures such as decoder-only models. The research areas span speech translation, text-to-speech tasks, and assistive technologies, providing a foundation for future work in audio-to-text translation systems aimed at improving accessibility for the hard of hearing.

Methods and materials

This study aims to develop an efficient framework for converting audio messages from popular messaging platforms into text, focusing on enhancing accessibility for hearing-impaired individuals. The methodology is structured into three main phases: Information Collection, Preprocessing, and Conversion of Audio Files to Text. Each phase plays a critical role in ensuring the accuracy and reliability of the final output, enabling effective communication for users with hearing impairments [22]. Figure 1 presents the architecture of the model proposed in this study.

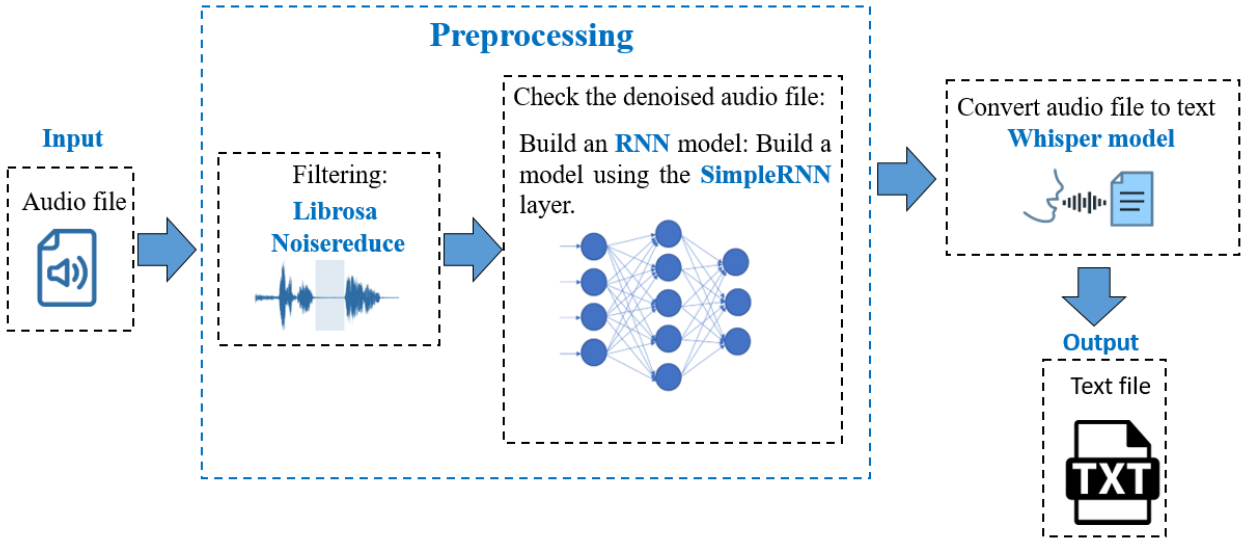

Figure 1. Architecture of the proposed model.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

Information Collection

In this section, we detail the collection of audio files from popular messaging platforms in Kazakhstan, including WhatsApp, Telegram, and Viber. A total of 150 audio files were collected from various group and individual chats to ensure a diverse dataset representing different accents, dialects, and speaking styles prevalent in the Kazakh language. The audio files ranged in duration from 10 seconds to 2 minutes, with an average duration of approximately 45 seconds. The dataset included a mix of audio recordings with varying levels of background noise, such as wind, street sounds, and indoor conversations, to simulate real-world conditions. These characteristics ensured that the dataset was representative of typical audio messaging scenarios encountered by hearing-impaired individuals.

Preprocessing

The pre-processing step is to improve the clarity of audio files by reducing background noise while preserving the integrity of human speech. This ensures that the Whisper-style transcription is as accurate and reliable as possible. The process involves a series of steps using specialized tools including Librosa, Noisereduce and SimpleRNN (recurrent neural network) model.

The first step involves downloading audio files and extracting important features with Librosa. Librosa is a widely used Python library for audio signal analysis. It facilitates various tasks such as loading audio, converting it to digital formats and extracting basic features such as frequency, amplitude and waveform characteristics. This allows us to present the audio in a format suitable for further processing [23].

Noise reduction is then applied to remove background noise. Noisereduce is a Python package that uses spectral gating, a noise reduction technique, to filter out unwanted sounds while preserving the clarity of the human voice. By using this method, significantly reduce the interference caused by background noises such as background conversations, street sounds or other environmental disturbances [24]. This step is essential to ensure that the underlying speech signal remains intact for subsequent transcription.

Use the SimpleRNN model to evaluate the effectiveness of the denoising process. SimpleRNN, a type of recurrent neural network, is designed to process data sequences such as audio signals, taking temporal dependencies into account. In this context, the SimpleRNN model is trained to evaluate the quality of processed audio files [25]. It analyzes various factors, including speech clarity, signal-to-noise ratio, and preservation of human speech characteristics. The model detects potential distortions and verifies that the noise reduction process was successful in preserving a high-quality audio signal.

The combination of Librosa, Noisereduce and SimpleRNN provides a comprehensive pre-processing workflow that improves the quality of audio files. This process is key to preparing data for accurate transcription with the Whisper model.

Whisper Model

The final step in the workflow involves converting the preprocessed audio files into text using **Whisper**, a machine learning model for speech recognition and transcription, created by OpenAI and first released as open-source software in September 2022. Whisper is a state-of-the-art automatic speech recognition (ASR) system that leverages deep learning techniques to perform accurate and robust transcription of audio, even in challenging environments with background noise or non-standard accents [26]. This section describes how the model is employed to transcribe the audio files gathered in our study.

Once the audio files have undergone noise reduction and quality verification through SimpleRNN, they are ready for conversion into text. The Whisper model is well-suited for this task due to its multi-lingual capabilities and its robustness in handling variations in audio quality. Whisper's architecture is based on a transformer neural network, making it capable of captur-

DOI: 10.37943/22SNOK5872

ing long-term dependencies in audio sequences, which is particularly important for recognizing continuous speech patterns and preserving context over long utterances.

We utilize the base Whisper model, which is pre-trained on a large and diverse dataset of audio samples, including multiple languages, making it ideal for handling the Kazakh language in this study. The model processes audio input in various formats, including .wav, .mp3, and .ogg, commonly found in messenger applications like WhatsApp and Telegram. The transcription process is performed in two main stages:

1. ***Feature Extraction***: The Whisper model first converts the audio signals into mel-spectrograms, a time-frequency representation of sound. This transformation breaks down the audio into a grid of frequency components over time, essential for detecting and understanding the nuances of speech, such as tone, pitch, and pauses. This step allows the model to recognize phonemes, words, and sentences, even from noisy or distorted audio.
2. ***Transcription****:* Once the audio is represented as a mel-spectrogram, the Whisper model processes it through its transformer-based architecture. The model segments the audio into overlapping chunks, each analyzed in relation to the preceding and following segments to maintain context. This design helps Whisper handle overlapping speech or rapid transitions in conversational dialogues. The output is a transcription of the speech in plain text format. The Whisper model is configured to transcribe directly into the Kazakh language, which is particularly important for our dataset of Kazakh audio files collected from popular messengers.

Throughout the transcription process, Whisper also offers the ability to detect pauses and emphasize changes in tone or sentiment, valuable for conversational analysis. Additionally, it provides timestamped outputs, enabling the tracking of when specific phrases or words were spoken, which can be used for further analysis of conversation dynamics in future research.

Input Representation

The model processes raw audio inputs by converting them into log-Mel spectrograms, whichserve as the primary input representation. This transformation involves the following operations (1):

$$S(f,t) = \log\left(\sum_k |X(f,k)|^2 * W(f,k)\right) \quad (1)$$

where $S(f, t)$ denotes the spectrogram intensity at frequency f and time t, $X(f, k)$ is the Short-Time Fourier Transform (STFT) of the input signal, and $W(f, k)$ represents the Mel filter bank. This spectral representation captures the time-frequency characteristics of the audio input, optimized for downstream ASR tasks.

Encoder-Decoder Transformer Architecture

The Whisper model employs a sequence-to-sequence transformer framework consisting of an encoder and a decoder.

Encoder

The encoder maps the input spectrogram x into a sequence of latent representations h^1, where l denotes the layer index (2):

$$h^l = TransformerLayet\left(h^{l-1}\right), h^0 = Embedding(x) \quad (2)$$

Each transformer layer incorporates multi-head self-attention and a feedforward network, expressed as follows (3):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

where Q, K and V represent the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors (4):

$$FFN(x) = W_2 * ReLU(W_1 * x + b_1) + b_2 \quad (4)$$

where W_1, W_2, b_1, and b_2 are learnable parameters.

Residual connections and layer normalization are applied to stabilize training (5):

$$h^l = LayerNorm(h^{l-1} + FFN(Attention(h^{l-1}))) \quad (5)$$

Decoder

The decoder generates the transcription as a sequence of tokens $y - (y_1, y_1 \ldots ., y_T)$. At each step ttt, the decoder outputs a probability distribution over the vocabulary (6):

$$P(y_t|y < t, x) = Softmax(W * h_t^{decoder} + b) \quad (6)$$

where $h_t^{decoder}$ is computed using cross-attention with encoder outputs (7):

$$h_t^{decoder} = TransformerLayer(h_{t-1}^{decoder}, h^{encoder}) \quad (7)$$

Training Objective

The model is trained to minimize the cross-entropy loss, which measures the divergence between predicted and ground truth token distributions (8):

$$L = -\sum_{t=1}^{T} \log P(y_t^*|y_{<t}^*, x) \quad (8)$$

where y_t^* is the ground truth token at step t, and T is the total number of tokens in the sequence.

Multilingual and Multitask Capability

Whisper incorporates language-specific tokens and task directives, enabling multilingual transcription and multitask learning. The tokenization process ensures flexibility in adapting the model to various linguistic and functional domains.

Inference and Decoding

During inference, the model employs beam search to maximize the posterior probability of the transcription sequence (9):

$$Y^* = arg\max_Y P(Y|x) \quad (9)$$

where Y represents the set of possible token sequences. Beam search considers multiple hypotheses at each decoding step to improve the accuracy of the generated text.

Whisper is instrumental in accurately converting the audio files into text, even when working with noisy or low-quality recordings. Its ability to handle complex acoustic conditions and multiple languages makes it a suitable choice for the Kazakh-language audio data in this study. The transcribed text is subsequently used for analysis and evaluation of the Whisper model's performance in assisting hearing-impaired individuals by transforming audio messages into readable text.

System overview

In this section, first, one audio message received by Whatsapp messenger is denoised using Librosa library and Noisereduce python package, and then the sound quality of the audio file

DOI: 10.37943/22SNOK5872

is checked using SimpleRNN. Once the audio file is ready to be converted to text, the final step is converting the audio to text using the Whisper model.

Denoise an audio file

This research aims to improve the quality of audio recordings by effectively reducing background noise. To achieve this, several libraries adapted to sound processing and noise reduction were used, namely:

```
import librosa
import noisereduce as nr
import soundfile as sf
from pydub import AudioSegment
import os
```

The main goal was to de-noise one person's message from WhatsApp messenger. This recording had significant background wind noise, which made it difficult to clearly recognize speech. The process involved the following steps:

Audio download: The audio file is loaded using the Librosa library, which allows you to extract the relevant features needed for editing.

Noise reduction: Using the noise reduction library, we applied spectral gating to effectively filter out wind noise while preserving the integrity of spoken words. This method amplifies the speech signal and reduces interference from ambient sounds.

Create an output: The edited audio file is saved using the SoundFile library, which can be accessed and used for subsequent transcription.

Results

The result of this muting process was a significantly improved audio recording with significantly reduced background wind noise. This improvement is very important because it allows for accurate transcription of spoken content in the later stages of the study.

After successfully loading the audio, we applied the Noisereduce library to perform spectral gating, effectively filtering out the background wind noise while preserving the speech signal. The denoised audio was subsequently saved for further analysis.

To evaluate the quality of the processed audio, we employed a SimpleRNN model. The training process consisted of 10 epochs, during which the model iteratively learned to assess audio quality based on the features extracted from the denoised file. The training results, represented by loss values, were recorded as follows:

Epoch 1/10: loss: 4.4859e-04
Epoch 2/10: loss: 4.2829e-04
Epoch 3/10: loss: 4.3410e-04
Epoch 4/10: loss: 4.2496e-04
Epoch 5/10: loss: 4.1832e-04
Epoch 6/10: loss: 4.1700e-04
Epoch 7/10: loss: 4.1882e-04
Epoch 8/10: loss: 4.2082e-04
Epoch 9/10: loss: 4.2103e-04
Epoch 10/10: loss: 4.1979e-04

The progressive decrease in loss values across the training epochs indicates that the SimpleRNN model effectively learned to evaluate the audio quality, confirming the success of the noise reduction process. These results demonstrate that the speech clarity was maintained, which is critical for ensuring accurate transcription in the following steps of our research. Displayed in Figure 3 are the amplitude graphs corresponding to the audio files.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

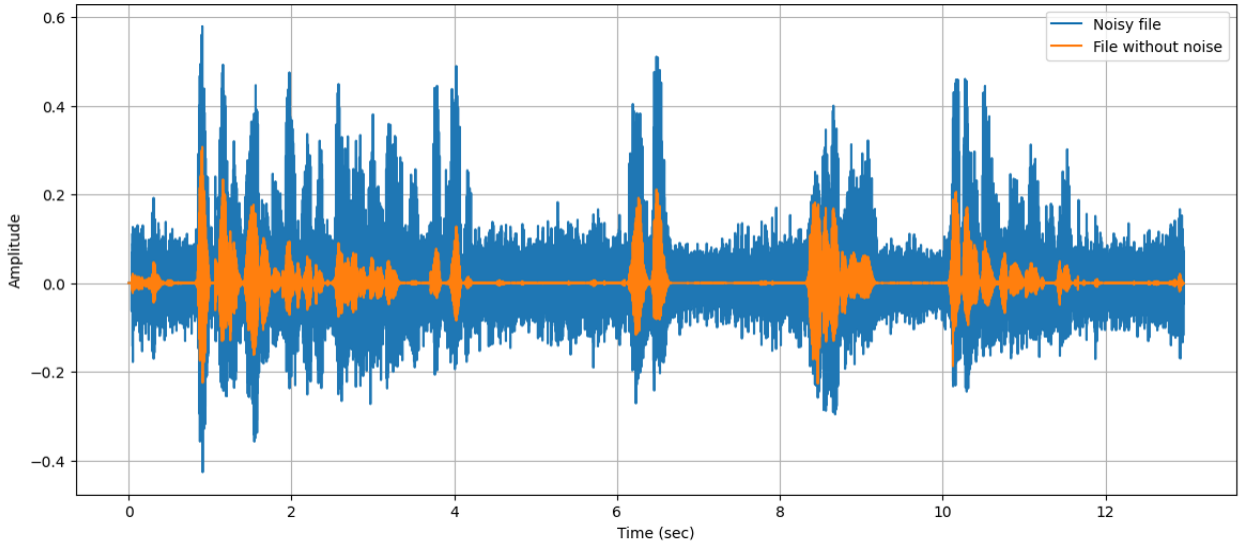

Figure 3. Amplitude graphs of Audio files

To further evaluate the effectiveness of the denoising process, we measured the signal-to-noise ratio (SNR) and word error rate (WER) before and after applying the Noisereduce package. The average SNR improved from 8.2 dB in the original audio files to 15.6 dB in the denoised files, indicating a significant reduction in background noise while preserving speech clarity. Additionally, we assessed the transcription accuracy by comparing the word error rate (WER) of the Whisper-Large model on the original and denoised audio files. The WER decreased from 18.5% for the original noisy audio to 6.3% for the denoised audio, demonstrating a substantial improvement in transcription accuracy due to the enhanced audio quality. These quantitative metrics, combined with the progressive decrease in SimpleRNN loss values (from 4.4859e-04 in Epoch 1 to 4.1979e-04 in Epoch 10), confirm the success of the denoising process in preparing the audio for accurate transcription.

Audio File Conversion to Text

In the subsequent phase of the research, the audio file was converted to text using the Whisper module, a powerful tool for speech recognition and transcription developed by OpenAI. The process began with the successful saving of the cleaned audio file, which confirmed that the audio had been adequately prepared for transcription. The following code snippet shows the process of converting an audio file to text using the Whisper module:

```
uploaded = files.upload()
model = whisper.load_model("large")

def transcribe_audio(file_path):
    audio = whisper.load_audio(file_path)
    audio = whisper.pad_or_trim(audio)
    result = model.transcribe(audio, language="kk")
    print("Transcript: ", result["text"])

audio_file = next(iter(uploaded))
     transcribe_audio(audio_file)
```

DOI: 10.37943/22SNOK5872

During this conversion, several warnings were generated that warrant attention. One warning highlighted the use of the torch.load function with the parameter weights_only=False. This setting poses potential security risks when loading untrusted model data, as it allows for the execution of arbitrary code during unpickling. The recommendation is to set weights_only=True in future applications to ensure a safer loading process.

Additionally, a warning regarding the use of floating-point precision was issued, indicating that FP16 (16-bit floating point) is not supported on the CPU. As a result, the system defaulted to using FP32 (32-bit floating point) for the transcription task. This is an important consideration for ensuring compatibility and optimizing performance during audio processing.

To evaluate the effectiveness of various Whisper model sizes in converting Kazakh audio to text, we conducted experiments using four versions: Whisper-Tiny, Whisper-Small, Whisper-Medium, and Whisper-Large. Each model was tested on the same audio file to compare transcription accuracy and overall performance.

The audio file used for these experiments contained a complex scenario with background noise and human speech, which is typical in real-world settings where hearing-impaired individuals might benefit from audio-to-text translations. Below are the results of the transcription outputs generated by each model.

Table 2. Transcription Results from Different Whisper Model Sizes

Model	Transcription
Tiny	Тәжірибе үшін жазып отырғаны шындық. Ауылда дауылды дауылды да, адамдар да болды.
Small	Ол тәжіргей үшін жызып жатқан мүнкін факт болып жатыр. Айылда шуы және адам дауыл жацлап болды.
Medium	Бұл тәжіргей үшін жазылып жатқан әлде файлы болып жатыр. Айылда шуы және адам дауысымен жазылып жатыр.
Large	Бұл тәжірбие үшін жазылып жатқан файл болып табылады. Бұл файлда шу мен адам дауысымен жазылып жатыр.

As observed in Table 2, the Whisper-Tiny model produced an output that was intelligible but exhibited significant errors, particularly in sentence structure and the interpretation of background noise. The Whisper-Small model improved slightly, but still showed inconsistencies in word choice and transcription of specific sounds.

The Whisper-Medium model demonstrated better accuracy, with more coherent sentence structures and fewer misinterpretations, although it still struggled with some nuances of human speech and noise differentiation. The Whisper-Large model, however, outperformed the other versions, delivering the most accurate transcription with minimal errors. It successfully distinguished between human speech and background noise, providing a highly readable and accurate transcription.

These results suggest that larger Whisper models, while requiring more computational resources, offer superior transcription accuracy, especially in noisy environments. For applications focused on improving accessibility for hearing-impaired individuals, where accuracy is paramount, the Whisper-Large model appears to be the most effective solution.

Discussion

The Whisper model used in this study demonstrated exceptional versatility and the ability to deliver precise results. It achieved high accuracy in converting audio recordings to text, outperforming several existing solutions with distinct advantages. While the literature review

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

identified algorithms and models such as PromptTTS, mSLAM, Fairseq S2T, and SLM that performed well in specific tasks, each of these approaches had notable limitations.

PromptTTSprimarily excelled in speech synthesis but was not well-suited for converting audio recordings to text [12]. Similarly, the mSLAMmodel proved effective in multimodal translation tasks but faced challenges in handling audio files with significant noise levels [13]. Although Fairseq S2T delivered promising results in converting audio to text, its performance diminished when dealing with noisy audio recordings [10]. On the other hand, the SLM model [19], based on statistical approaches, was considerably less capable than modern neural network-based methods.

In this context, the Whisper model offered several distinct advantages. First, it demonstrated an exceptional ability to accurately recognize information from audio files, even in cases with high noise levels. While tools like Librosa and Noisereduce were used for noise reduction in this study, Whisper effectively produced reliable results without requiring extensive pre-processing. Second, the model stood out for its ability to recognize multilingual audio recordings, which was particularly significant for audio files in the Kazakh language.

Another key advantage of Whisper is its neural network-based architecture, which showcased superior performance in processing large datasets and understanding diverse linguistic contexts. Furthermore, the model's resilience to noise and its data processing speed enhanced the quality of the study's outcomes.

Conclusıon

This study aimed to explore the efficacy of the Whisper model for audio-to-text transcription, particularly for enhancing accessibility for hearing-impaired individuals. The process began with the collection of audio recordings from WhatsApp messenger, which were subjected to a comprehensive denoising procedure using the Librosa library and Noisereduce package. By applying spectral gating techniques, we successfully reduced significant background noise, particularly wind interference, while preserving the clarity of the spoken content. The qualitative assessment of the processed audio was performed using a SimpleRNN model, which demonstrated a consistent decrease in loss values across training epochs, indicating that the audio quality had been effectively enhanced.

Following the successful denoising, the cleaned audio files were transcribed using the Whisper model, renowned for its robust capabilities in speech recognition. The transcription process yielded accurate text output in Kazakh, affirming the model's effectiveness even in challenging audio environments with background noise. However, it is essential to note the warnings encountered during the transcription phase, particularly regarding the security implications of loading untrusted model data and the use of appropriate floating-point precision settings to ensure optimal performance.

The successful implementation of these methodologies highlights the potential of machine learning technologies in facilitating better communication and information access for individuals with hearing impairments. The findings underscore the significance of high-quality audio input, not only for accurate transcription but also for enhancing the overall user experience. Future research could expand upon these results by investigating additional noise reduction strategies and exploring the Whisper model's applicability across various languages and dialects, ultimately contributing to a more inclusive digital environment.

This study not only demonstrates the practical applications of advanced machine learning techniques for audio transcription but also emphasizes the critical role of audio quality in achieving reliable outcomes, paving the way for future advancements in accessibility technologies.

References

[1] International Health Organization. (2024, September 26). Deafness and hearing loss. https://www.who.int/ru/news-room/fact-sheets/detail/deafness-and-hearing-loss
[2] Martini, A., Cozza, A., & Di Pasquale Fiasca, V.M. (2024). The inheritance of hearing loss and deafness: A historical perspective. *Audiology Research*, *14*(1), 116-128. https://doi.org/10.3390/audiolres14010012
[3] Lan, S., Ye, L., & Zhang, K. (2023). Applying mmWave Radar Sensors to Vocabulary-Level Dynamic Chinese Sign Language Recognition for the Community With Deafness and Hearing Loss. *IEEE Sensors Journal*. https://doi.org/10.1109/JSEN.2023.3241237
[4] Kral, A., & Sharma, A. (2023). Crossmodal plasticity in hearing loss. *Trends in neurosciences*, *46*(5), 377-393. https://doi.org/10.1016/j.tins.2023.02.004
[5] Baballe, M. A., Garba, A., & Dahiru, M. (2023). Reasons for Deafness and Hearing Loss. *Available at SSRN 4629219*. https://ssrn.com/abstract=4629219
[6] Podury, A., Jiam, N.T., Kim, M., Donnenfield, J.I., & Dhand, A. (2023). Hearing and sociality: the implications of hearing loss on social life. *Frontiers in Neuroscience*, *17*, 1245434. https://doi.org/10.3389/fnins.2023.1245434
[7] de Guimaraes, T.A.C., Arram, E., Shakarchi, A.F., Georgiou, M., & Michaelides, M. (2023). Inherited causes of combined vision and hearing loss: clinical features and molecular genetics. *British Journal of Ophthalmology*, *107*(10), 1403-1414. https://doi.org/10.1136/bjo-2022-322062
[8] Jiang, L., Wang, D., He, Y., & Shu, Y. (2023). Advances in gene therapy hold promise for treating hereditary hearing loss. *Molecular Therapy*, *31*(4), 934-950. https://doi.org/10.1016/j.ymthe.2023.01.022
[9] Mohammed, H. B., & Cavus, N. (2024). Utilization of Detection of Non-Speech Sound for Sustainable Quality of Life for Deaf and Hearing-Impaired People: A Systematic Literature Review. *Sustainability*, *16*(20), 8976. https://doi.org/10.3390/su16208976
[10] Wang, C., Tang, Y., Ma, X., Wu, A., Popuri, S., Okhonko, D., & Pino, J. (2020). Fairseq S2T: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*. https://arxiv.org/abs/2010.05171
[11] Xu, C., Ye, R., Dong, Q., Zhao, C., Ko, T., Wang, M., ... & Zhu, J. (2023). Recent advances in direct speech-to-text translation. *arXiv preprint arXiv:2306.11646*. https://arxiv.org/abs/2306.11646
[12] Guo, Z., Leng, Y., Wu, Y., Zhao, S., & Tan, X. (2023, June). Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICASSP49357.2023.10094829
[13] Bapna, A., Cherry, C., Zhang, Y., Jia, Y., Johnson, M., Cheng, Y., ... & Conneau, A. (2022). mslam: Massively multilingual joint pre-training for speech and text. *arXiv preprint arXiv:2202.01374*. https://arxiv.org/abs/2202.01374
[14] Liu, Y., Zhu, J., Zhang, J., & Zong, C. (2020). Bridging the modality gap for speech-to-text translation. *arXiv preprint arXiv:2010.14920*. https://arxiv.org/abs/2010.14920
[15] Tang, Y., Pino, J., Wang, C., Ma, X., & Genzel, D. (2021, June). A general multi-task learning framework to leverage text data for speech to text tasks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6209-6213). IEEE. https://doi.org/10.1109/ICASSP39728.2021.9413686
[16] Matre, M. E., & Cameron, D. L. (2024). 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. https://doi.org/10.1080/17483107.2023.2243206
[17] Wu, J., Gaur, Y., Chen, Z., Zhou, L., Zhu, Y., Wang, T., ... & Wu, Y. (2023, December). On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 1-8). IEEE. https://doi.org/10.1109/ASRU59171.2023.10383163
[18] Higuchi, Y., Chen, N., Fujita, Y., Inaguma, H., Komatsu, T., Lee, J., ... & Watanabe, S. (2021, December). A comparative study on non-autoregressive modelings for speech-to-text generation. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 47-54). IEEE. https://doi.org/10.1109/ASRU51503.2021.9688364

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

[19] Wang, M., Han, W., Shafran, I., Wu, Z., Chiu, C.C., Cao, Y., ... & Wu, Y. (2023, December). Slm: Bridge the thin gap between speech and text foundation models. In*2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 1-8). IEEE. https://doi.org/10.1109/ASRU59171.2023.10383153

[20] Bhandari, A., Shah, S. B., Thapa, S., Naseem, U., & Nasim, M. (2023). Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1994-2003). https://doi.org/10.1109/CVPR52729.2023.00209

[21] Bano, S., Jithendra, P., Niharika, G. L., & Sikhi, Y. (2020, November). Speech to text translation enabling multilingualism. In *2020 IEEE International Conference for Innovation in Technology (INOCON)* (pp. 1-4). IEEE. https://doi.org/10.1109/INOCON50539.2020.9298288

[22] Golla, Ramsri Goutham (2024-09-26). "Here Are Six Practical Use Cases for the New Whisper API". *Slator*. Archived from the original on 2024-09-26. Retrieved 2024-09-26. https://slator.com/here-are-six-practical-use-cases-for-the-new-whisper-api

[23] Albahra, S., Gorbett, T., Robertson, S., D'Aleo, G., Kumar, S.V.S., Ockunzzi, S.,... & Rashidi, H.H. (2023, March). Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts. In *Seminars in Diagnostic Pathology* (Vol. 40, No. 2, pp. 71-87). WB Saunders. https://doi.org/10.1053/j.semdp.2023.02.002

[24] Samadi, M.E., Mirzaieazar, H., Mitsos, A., & Schuppert, A. (2024). Noisecut: a python package for noise-tolerant classification of binary data using prior knowledge integration and max-cut solutions. *BMC bioinformatics*, *25*(1), 155. https://doi.org/10.1186/s12859-024-05693-1

[25] Gholami, H., Mohammadifar, A., Golzari, S., Song, Y., & Pradhan, B. (2023). Interpretability of simple RNN and GRU deep learning models used to map land susceptibility to gully erosion. *Science of the Total Environment*, *904*, 166960. https://doi.org/10.1016/j.scitotenv.2023.166960

[26] Zezario, R.E., Chen, Y.W., Fu, S.W., Tsao, Y., Wang, H.M., & Fuh, C.S. (2024, July). A study on incorporating Whisper for robust speech assessment. In *2024 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICME52920.2024.10462847