MODELLING AND DATA SCIENCE - PROFESSIONAL DESIGN PROJECT FINAL REPORT 2024/25

PROJECT TITLE   
Transcription, translation and text-to-speech of the voice from short audio and video files using artificial intelligence methods

**Authors***Marta Haik***,** *Ansar Shilibekov****,*** *Filip Chmiel****,*** *Radosław Lech****,*** *Katarzyna Borsiak****,*** *Zuzanna Gielec****,***

*Adrian Florczak****,*** *Juliusz Ciążyński****,*** *Hubert Wodziak*

Lodz University of Technology, Lodz, Poland

**Supervisor:**

*PhD Eng. Paulina Komar*  
Lodz University of Technology, Lodz, Poland

Date of submission: XX.01.2025

ABSTRACT This template provides instructions for your final report. The abstract must be a concise yet comprehensive reflection of what your project is about. In particular, the abstract must be self-contained, without abbreviations, footnotes, or references. It should be a microcosm of the full report, describing the problem, solution, technical summary of the prototype and how it was tested.

The abstract should be below 1000 characters. Ensure that your abstract reads well and is grammatically correct.

KEYWORDS Enter a list of keywords or phrases in alphabetical order, separated by commas.

I. INTRODUCTION

Our project is focused on developing a solution for audio-to-text and text-to-audio translation conversion using deep learning models. The report aims to investigate speech recognition and translation accuracy by comparing team-made models with existing ones.

1. **PURPOSE AND OBJECTIVES OF THE PROJECT**

*The primary goal of the project is to design a system capable of converting audio files into text efficiently and accurately. The problem arises from the growing need for automated transcription tools, particularly in applications such as meeting notes, interviews, and content creation. The scope of this project focuses on building a system that outperforms existing solutions by training two pre-existing models and comparing their performance to a transformer-based model. The objective is to evaluate which model provides the best accuracy and speed for transcription tasks.*

*B. PROBLEM FINDING*

*The purpose of the project is to develop a functional solution that can convert audio input into text with a high level of accuracy. The expected result is a system that can process audio files and output the transcribed text in a matter of seconds, with minimal errors. The problem formulation is narrowed down by considering the limitations of existing models, particularly their inability to handle certain accents, background noise, and diverse speech patterns. The research conducted aims to identify which model can overcome these issues effectively.*

II.  IDEA FINDING

This section will describe the process of finding your solution.

*A. STATE OF THE ART*

*Existing speech recognition technologies are primarily based on deep learning models such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks. These models have been successful in many applications, but they are limited by their inability to handle noise and accents effectively. Recently, transformer-based models like BERT and GPT have shown significant promise in various natural language processing tasks, including speech-to-text conversion. Some commercial solutions, such as Google Speech-to-Text and IBM Watson, have set high standards, but they are not perfect and can be expensive for large-scale usage.*

*For reference, studies like [1] have shown that while RNNs work well for smaller datasets, transformer-based models outperform them on larger, more complex datasets. Current research focuses on improving these models for better adaptability to various audio qualities and languages.*

*B. INNOVATIVE IDEAS*

*Our team aims to compare the performance of two pre-existing models (RNN and CNN) with a transformer-based model for the speech-to-text task. We believe that the transformer model will provide better accuracy due to its ability to process long-range dependencies in speech patterns. The innovative aspect of our solution lies in the comparison of these models under real-world conditions, such as noisy audio environments and diverse speech accents.*

*C. MAIN IDEA SELECTION AND JUSTIFICATION*

*After testing both pre-existing models (RNN and CNN) and transformer-based models, we concluded that the transformer model has the potential to achieve the best results due to its architecture, which allows it to handle complex data more effectively. We chose Python as the programming language for its rich ecosystem of libraries such as TensorFlow and PyTorch, which support the implementation of both pre-existing and transformer-based models. We will also use the pre-existing models and tutorials available online as a foundation for our project, saving development time and effort.*

III. SOLUTION IMPLEMENTATION  
A. ORGANIZATION AND MANAGEMENT

The project was divided into 3 main phases:

Phase 1: Research. We began by exploring existing models and comparing their features. Small description.

Phase 2: Model training and testing. …

Phase 3: Evaluation and comparison. We conducted tests to compare the accuracy, speed, and adaptability of the models under various conditions, such as noisy and accented speech.

Weekly scrum meetings were held to ensure effective communication, review progress, and tackle challenges. We used a sprint-based approach to keep track of tasks and milestones.

B. TECHNICAL DETAILS

The solution integrates two pre-existing models (RNN and CNN) for speech recognition and compares them against a transformer-based model.

1. speech to text

2. transformer

3. text to speech

C. COMPARISON WITH EXISTING SOLUTION

In this section, we analyze an existing solution to speech-to-speech translation from the GitHub repository [Realtime-Speech-to-Speech-Translation](https://github.com/kensonhui/Realtime-Speech-to-Speech-Translation) [ref]. This solution offers real-time speech translation, converting spoken language into another language using a combination of deep learning techniques. We will evaluate its architecture, performance, and limitations in comparison to the approach we are implementing in this project.

IV. VERIFICATION

To verify the effectiveness of the solution, we conducted several tests:

* Performance metrics: word error rate (WER) and the time it took for each model to process the audio,
* Key scenarios: scenarios such as clear speech, noisy environments, diverse accents,
* Validation: comparing the models' transcriptions with a manually created translation,
* Plots of WER vs. time for each model/ accuracy plots across different scenarios (helps to illustrate the performance differences)

**V. LIFE CYCLE ANALYSIS.**

The LCA consists of the materials, energy consumption, and environmental impacts involved throughout the lifecycle of the speech-to-speech translation system:

Short comments on each stage:

A. Hardware requirements:

computational power, energy consumption, data storage

B. Software development:

training data, model training, deployment

C. Impact assessment methods

energy consumption,

resources usage

D. Interpretation of results

recommendations

limitations

**VI. LEGAL ASPECTS AND PROTECTION OF INTELLECTUAL PROPERTIES**

The project uses existing models and public datasets for training, which are subject to the terms of their respective licenses.

We did include a checkbox with the link to the following clause that the user must agree to before running the program [ref]:

\*the copyright/terms of use snippet\*

As can be seen above, the clause states that users are responsible for ensuring that the audio files they input into the system belong to them or can be legally used. Any generated code or modifications to existing models will be made available under an open-source license, to ensure the intellectual property is protected and freely accessible for educational and research purposes.

VII. CONCLUSIONS AND PERSPECTIVES

Summarize your work, highlighting strengths and weaknesses of its results. How do you see a potential follow-up of the project?

APPENDIX

Links to code repositories online.

ACKNOWLEDGMENTS

We’d like to thank our supervisor, PhD Eng. Paulina Komar, for all her help throughout this project. Her support, patience, and encouragement kept us motivated, even when things got disorganized or tough.

**REFERENCES**

1. A. Author, “Paper title,” *Journal Name*., vol. 13, no. 1, pp. 11-23, Jan. 1995. DOI: 10.1109/TTHZ.2016.2544142
2. B. Author, “Paper title,” *Journal Name*., vol. 13, no. 1, pp. 11-23, Jan. 1995. DOI: 10.1109/TTHZ.2016.2544142

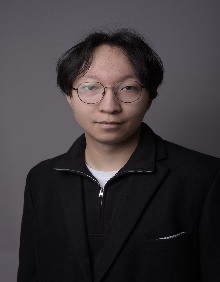
FIGURE 1.  Prepare good resolution of images (at least 300dpi, preferably 600dpi) and max 88mm wide. If the figure comes from external reference, cite it here [2]

AUTHOR BIOS AND CONTRIBUTIONS

Obraz zawierający Ludzka twarz, osoba, ubrania, okulary/szklanki

Opis wygenerowany automatycznieMARTA K. HAIK was born in Lodz in 1999 and since 2020 has been enrolled at University XXX, majoring in XXX. If he/she has some job experience you can put it in here as well.

The second paragraph should summarize the author’s contributions to the project. Was he/she primarily involved in research? Prototype design? Coding? Managed/organized meetings? Prepared what sections of the report? You can write out longer description of what the author did, spreading to below the photo. The description should be between 500 and 1000 characters long. The photos should be 3,26 cm x 2,56 cm and either all colored or all greyscale.

ANSAR SHILIBEKOV was born in Lodz in 1999 and since 2020 has been enrolled at University XXX, majoring in XXX. If he/she has some job experience you can put it in here as well.

The second paragraph should summarize the author’s contributions to the project. Was he/she primarily involved in research? Prototype design? Coding? Managed/organized meetings? Prepared what sections of the report? You can write out longer description of what the author did, spreading to below the photo. The description should be between 500 and 1000 characters long. The photos should be 3,26 cm x 2,56 cm and either all colored or all greyscale.

**Obraz zawierający Ludzka twarz, osoba, uśmiech, Warga

Opis wygenerowany automatycznieFILIP CHMIEL** was born in Lodz in 1999 and since 2020 has been enrolled at University XXX, majoring in XXX. If he/she has some job experience you can put it in here as well.

The second paragraph should summarize the author’s contributions to the project. Was he/she primarily involved in research? Prototype design? Coding? Managed/organized meetings? Prepared what sections of the report? You can write out longer description of what the author did, spreading to below the photo. The description should be between 500 and 1000 characters long. The photos should be 3,26 cm x 2,56 cm and either all colored or all greyscale.

Obraz zawierający Ludzka twarz, osoba, uśmiech, Warga

Opis wygenerowany automatycznieRADOSŁAW LECH was born in Lodz in 1999 and since 2020 has been enrolled at University XXX, majoring in XXX. If he/she has some job experience you can put it in here as well.

The second paragraph should summarize the author’s contributions to the project. Was he/she primarily involved in research? Prototype design? Coding? Managed/organized meetings? Prepared what sections of the report? You can write out longer description of what the author did, spreading to below the photo. The description should be between 500 and 1000 characters long. The photos should be 3,26 cm x 2,56 cm and either all colored or all greyscale.

Obraz zawierający Ludzka twarz, osoba, uśmiech, Warga

Opis wygenerowany automatycznieKATARZYNA BORSIAK was born in Lodz in 1999 and since 2020 has been enrolled at University XXX, majoring in XXX. If he/she has some job experience you can put it in here as well.

The second paragraph should summarize the author’s contributions to the project. Was he/she primarily involved in research? Prototype design? Coding? Managed/organized meetings? Prepared what sections of the report? You can write out longer description of what the author did, spreading to below the photo. The description should be between 500 and 1000 characters long. The photos should be 3,26 cm x 2,56 cm and either all colored or all greyscale.

Obraz zawierający Ludzka twarz, osoba, uśmiech, Warga

Opis wygenerowany automatycznieZUZANNA GIELEC was born in Lodz in 1999 and since 2020 has been enrolled at University XXX, majoring in XXX. If he/she has some job experience you can put it in here as well.

The second paragraph should summarize the author’s contributions to the project. Was he/she primarily involved in research? Prototype design? Coding? Managed/organized meetings? Prepared what sections of the report? You can write out longer description of what the author did, spreading to below the photo. The description should be between 500 and 1000 characters long. The photos should be 3,26 cm x 2,56 cm and either all colored or all greyscale.

Obraz zawierający Ludzka twarz, osoba, uśmiech, Warga

Opis wygenerowany automatycznieADRIAN FLORCZAK was born in Lodz in 1999 and since 2020 has been enrolled at University XXX, majoring in XXX. If he/she has some job experience you can put it in here as well.

The second paragraph should summarize the author’s contributions to the project. Was he/she primarily involved in research? Prototype design? Coding? Managed/organized meetings? Prepared what sections of the report? You can write out longer description of what the author did, spreading to below the photo. The description should be between 500 and 1000 characters long. The photos should be 3,26 cm x 2,56 cm and either all colored or all greyscale.

Obraz zawierający Ludzka twarz, osoba, uśmiech, Warga

Opis wygenerowany automatycznieJULIUSZ CIĄŻYŃSKI was born in Lodz in 1999 and since 2020 has been enrolled at University XXX, majoring in XXX. If he/she has some job experience you can put it in here as well.

The second paragraph should summarize the author’s contributions to the project. Was he/she primarily involved in research? Prototype design? Coding? Managed/organized meetings? Prepared what sections of the report? You can write out longer description of what the author did, spreading to below the photo. The description should be between 500 and 1000 characters long. The photos should be 3,26 cm x 2,56 cm and either all colored or all greyscale.

Obraz zawierający Ludzka twarz, osoba, uśmiech, Warga

Opis wygenerowany automatycznieHUBERT WODZIAK was born in Lodz in 1999 and since 2020 has been enrolled at University XXX, majoring in XXX. If he/she has some job experience you can put it in here as well.

The second paragraph should summarize the author’s contributions to the project. Was he/she primarily involved in research? Prototype design? Coding? Managed/organized meetings? Prepared what sections of the report? You can write out longer description of what the author did, spreading to below the photo. The description should be between 500 and 1000 characters long. The photos should be 3,26 cm x 2,56 cm and either all colored or all greyscale.