Systematic Literature Review Report

1. Research Project

1.1 Title

A program that eliminates a specific phrase from an audio recording

1.2 Supervisor

Marcin Pazio (KSiS WETI)

1.3 Goals and short description

The aim of the project is to create a program that eliminates specific phrases

From a source file. A phrase can be defined in at least one of the following ways:

- Indicated in the source file
- Given in the text form
- Given in the form of an audio recording other than the recording from the source file

The program has to locate the given phrase and perform at least one of the following modifications to the recording:

- Mute or replace with a continuous tone
- Replace with a distorted phrase (e.g. replace the word “tomato” with the word “rodimop”)
- Replace with another phrase given as text or audio recording

2. Systematic Literature Review Plan

2.1 Goals and questions

1. What are the currently used methods for identifying and locating phrases in audio recordings?
2. What are the relative performance differences between leading speech recognition services?

3. How do different model sizes and architectures (such as Whisper tiny, base, small, medium, and large) affect transcription accuracy as measured by Word Error Rate (WER)?
4. What is the relationship between model size (number of parameters) and transcription performance across different speech recognition systems?
5. How does transcription accuracy vary across different languages and accents for multilingual models like MMS and Whisper?

2.2 Keywords

- Artificial Intelligence (AI)
- Speech Recognition (SR)
- Keyword Spotting (KWS)
- Automatic Speech Recognition (ASR)
- Audio Transcription
- Word Error Rate
- Cloud Transcription Service
- Text-to-Speech

2.3 Search strings

- "Speech Recognition" OR "Automatic Speech Recognition" OR "ASR"
- ("Audio Transcription" OR "Speech Recognition") AND ("Word Error Rate" OR "WER")
- ("Speech Recognition Models") AND ("Performance Metrics" OR "Accuracy")
- ("Cloud Transcription Services") AND ("Speech Recognition" OR "ASR")
- ("Multilingual Speech Recognition") AND ("Language Models" OR "Performance")
- ("Audio Transcription") AND ("Speech Recognition") AND ("Automatic Speech Recognition")
- ("Whisper Model" OR "MMS Model") AND ("Speech Recognition")
- ("Speech Recognition") AND ("Model Size" OR "Parameters") AND ("Performance")

2.4 Literature Databases

- IEEE Xplore
- ACM
- Scopus
- ScienceDirect

2.5 Inclusion criteria

Year 2021 – 2025

Language: English

Type: Article

Publication Stage: Final

Subject Area: Computer Science

2.6 Exclusion criteria

- Lack of access to full text of the article
- No clear methodology description (missing information about the model used, input data, testing conditions)
- No performance evaluation metrics – such as Word Error Rate (WER), Character Error Rate (CER), Accuracy, Precision
- Not related to speech recognition, audio transcription or phrase detection

2.7 Quality Criteria

1. C1: Word sample size (1 = <= 1000 words, 5 >= 5000 words)
2. C2: Publication Year (1 = 2021 or earlier, 5 = 2025)
3. C3: Methological Rigor and Transparency

C3: Whether the study clearly describes its process (includes details on models, datasets, etc.)

Final evaluation of paper is based on following formula = (C1 + C2 + 2 * C3) / 4

2.8 Data extraction

The following will be extracted:

- Speech recognition libraries/models
- Performance metrics (WER, CER, Model size, Memory usage)
- Language performance

2.9 SLR process

The steps for Systematic Literature Review:

Database Selection and Collecting Articles:

- Scopus
- ACM Digital Library
- IEEE Xplore

- ScienceDirect

Data Collection:

- Articles will be collected and stored in .csv or .bib with relevant data about the articles including the abstract, we will have 4 .csv and .bib files in total
- .csv files will be edited using www.convertcsv.com, or visual studio code for .bib which allows to read abstracts and quicky delete rows (articles)

First stage:

- Articles will be filtered based on their relevance using abstracts and titles
- Decisions will be performed to keep or discard the article in their own database.

Second stage:

- Articles will be read
- Articles will be evaluated using the 3 criterias (2.5 to 2.7), certain articles will be accepted and the data will be extracted from them.

3. Systematic Literature Review Results

3.1 Results in numbers

To first stage I accepted given number of articles:

- ACM Digital Library - 353
- ScienceDirect - 100
- IEEE Xplore – 210
- Scopus - 231

To second stage I accepted given number of articles:

- ACM Digital Library - 6
- ScienceDirect – 8
- Scopus - 12
- IEEE Xplore - 14

3.2 Articles selected for data extraction

- IEEE Xplore - Automatic and Multilingual Speech Recognition and Translation by using Google Cloud API

Quality = (3 + 4 + 8)/4 = 3.75

- IEEE Xplore - Evaluating Automatic Transcription Models Utilising Cloud Platforms
  Quality = (4 + 4 + 8)/4 = 4

- ACM Digital Library - Scaling speech technologies to 1000+ languages
  Quality = (5 + 4 + 10)/4 = 4.75

- ScienceDirect - Evaluation of the efficiency of state-of-the-art Speech Recognition engines
  Quality = (5 + 2 + 8)/4 = 3.75

- IEEE Xplore - A Study of Audio-to-Text Conversion Software Using Whispers Model
  Quality = (4 + 3 + 8)/4 = 3.75

- IEEE Xplore - WhisperSum: Unified Audio-to-Text Summarization
  Quality = (4 + 4 + 8)/4 = 4

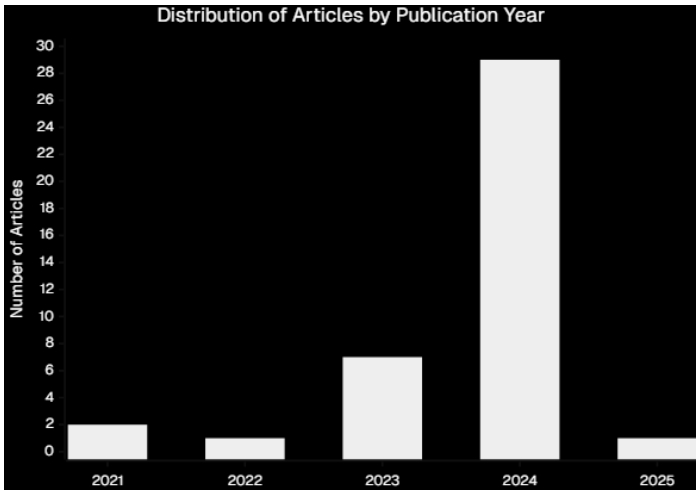
- IEEE Xplore - Speech Recognition Paradigms: A Comparative Evaluation of SpeechBrain, Whisper and Wav2Vec2 Models
  Quality = (3 + 4 + 10)/4 = 4.25

3.3 Snowballed articles

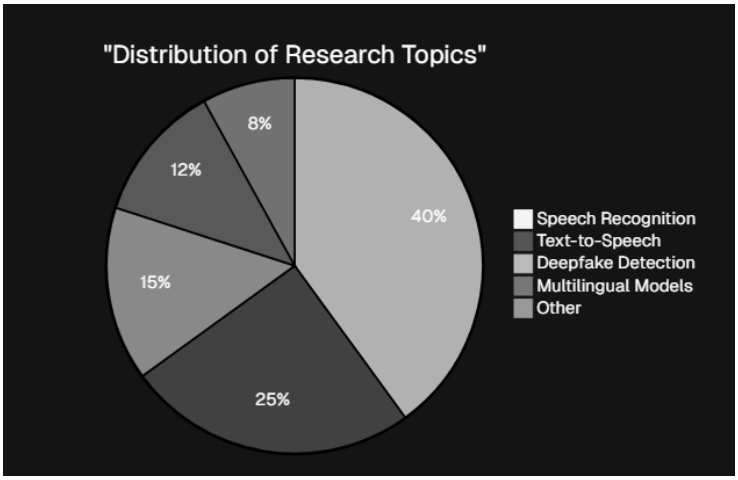
- “Wav2vec 2.0: A framework for Self-Supervised Learning of Speech Representations”: snowballed from “A Study of Audio-to-Text Conversion Software Using Whispers Model”
- “Is word error rate a good indicator for spoken language understanding accuracy”: snowballed from “Speech Recognition Paradigms: A Comparative Evaluation of SpeechBrain, Whisper and Wav2Vec2 Models”

3.4 Article Statistics

Distribution of Articles in Stage 2 by Publication Year



Distribution of Research Topics of Articles in Stage 2



3.5 Initial extracted data

Articles:

- Scaling speech technologies to 1000+ languages

Comparison of Speech Recognition Models

Model	Languages	Training Data (hours)	Model Size (M)	FLEURS-54 WER (test)
Whisper medium	99	680K	769M	50.1
Whisper large-v2	99	Not specified	1,550M	44.3
MMS	61	3K	965M	33.3
MMS + CC LM	61	3K	965M	20.7
MMS (LSAH)	61	3K	1,096M	31.0
MMS (LSAH) + CC LM	61	3K	1,096M	19.0
MMS	1,107	45K	965M	44.2
MMS + CC LM	1,107	45K	965M	24.8
MMS (LSAH)	1,107	45K	3,346M	32.5
MMS (LSAH) + CC LM	1,107	45K	3,346M	18.7

Word Errror Rate for different models for each specific language

	Whisper medium	Whisper large-v2	MMS L-61 noLM	MMS L-61 CC LM	MMS L-61 noLM LSAH	MMS L-61 CC LM LSAH	MMS L-1107 noLM	MMS L-1107 CC LM	MMS L-1107 noLM LSAH	MMS L-1107 CC LM LSAH
Amharic	229.3	140.3	48.7	30.7	52.4	32.5	52.9	30.1	53.3	31.1
Arabic	20.4	16.0	34.9	19.6	35.8	19.9	44.0	23.4	41.3	21.0
Assamese	102.3	106.2	29.5	18.8	28.4	18.6	37.6	21.2	30.5	19.2
Azerbaijani	33.1	23.4	40.7	21.3	38.3	19.8	45.0	21.2	40.1	19.1
Bengali	100.6	104.1	19.7	11.6	20.0	12.1	25.0	12.5	23.5	12.1
Bulgarian	21.4	14.6	23.4	13.1	23.9	13.3	27.9	12.9	25.5	13.5
Burmese	123.0	115.7	22.2	14.2	22.3	14.5	29.2	20.2	24.5	16.0
Catalan	9.6	7.3	18.1	11.0	18.1	11.0	25.9	11.5	20.1	10.8
Dutch	9.9	6.7	26.9	13.7	26.4	14.3	38.1	14.9	27.6	14.5
English	4.4	4.2	23.8	10.7	24.8	11.8	38.8	12.2	27.8	12.3
Filipino	19.1	13.8	19.3	11.9	19.4	12.2	26.2	13.5	20.2	12.4
Finnish	13.9	9.7	26.4	22.5	26.9	23.1	32.3	22.2	28.8	23.1
French	8.7	8.3	24.3	13.7	24.5	14.1	35.8	15.4	29.3	15.0
German	6.5	4.5	22.5	13.2	22.3	13.7	38.4	13.1	22.5	13.3
Greek	19.0	12.5	40.8	14.0	40.5	13.6	57.5	13.0	40.1	13.6
Gujarati	104.8	102.7	23.0	13.0	22.7	12.8	73.9	56.4	24.0	12.8
Hausa	106.6	88.9	35.9	26.7	36.3	27.3	40.4	26.7	38.3	26.4
Hebrew	33.1	27.1	68.5	44.8	66.6	41.5	78.7	50.9	67.1	40.0
Hindi	26.8	21.5	65.0	44.4	28.8	16.0	70.7	45.7	21.2	10.6
Hungarian	24.3	17.0	31.2	18.1	30.7	18.4	40.3	18.3	30.7	18.0
Icelandic	49.9	38.2	42.9	18.3	42.3	19.9	53.6	20.5	45.3	18.6
Indonesian	10.2	7.1	25.5	11.7	23.8	12.1	31.9	11.6	23.4	11.8
Javanese	67.9	68.5	32.8	19.6	32.8	20.0	58.8	27.2	34.2	19.5
Kannada	77.7	37.0	18.8	14.4	15.8	12.9	41.3	25.2	17.7	13.3
Kazakh	48.8	37.7	30.2	17.4	30.2	17.7	63.8	19.5	31.6	17.4
Khmer	103.8	128.9	26.0	19.9	25.7	19.8	70.7	52.4	26.7	19.7
Korean	16.4	14.3	58.7	37.5	59.9	37.3	82.1	58.2	68.3	40.1
Lao	101.4	101.5	48.9	45.4	24.2	22.8	62.1	56.6	22.6	16.9
Latvian	32.0	23.1	20.8	12.0	20.9	12.1	24.5	11.9	21.8	12.1
Malay	12.2	8.7	25.3	12.3	25.9	13.2	32.4	12.1	26.1	12.5
Malayalam	101.1	100.7	23.7	19.1	19.5	16.6	39.1	25.6	20.4	15.3
Marathi	63.2	38.3	32.5	19.0	19.2	13.5	28.0	14.9	20.9	13.4
Mongolian	103.7	110.5	55.7	29.3	54.9	32.9	67.7	28.7	55.3	32.3
Persian	41.0	32.9	39.7	22.9	39.9	22.5	44.4	21.3	42.9	22.0
Polish	8.0	5.4	21.5	11.4	20.8	11.6	33.0	11.0	25.1	11.3
Portuguese	5.0	4.3	16.1	10.8	16.3	10.8	19.3	10.2	17.7	10.5
Punjabi	102.0	102.4	41.4	29.9	30.4	20.7	99.0	91.0	31.0	19.8
Romanian	20.0	14.4	27.9	18.8	28.4	19.1	31.3	17.8	27.4	18.3
Russian	7.2	5.6	30.3	14.6	30.4	14.3	38.8	14.7	35.0	15.0
Shona	143.9	121.0	38.1	30.4	37.7	30.1	43.0	29.9	37.8	29.6
Somali	104.0	102.9	51.8	42.8	52.5	43.0	54.5	42.9	53.8	42.8
Spanish	3.6	3.0	12.2	7.8	12.4	8.2	14.0	7.8	14.0	8.7
Swahili	52.8	39.3	22.9	16.0	23.3	15.6	29.6	16.8	23.7	16.0
Swedish	11.2	8.5	29.9	17.4	30.5	17.5	38.2	17.2	33.5	17.4
Tajik	74.0	85.8	59.8	46.6	33.9	19.2	59.0	39.5	25.7	15.7
Tamil	23.1	17.5	24.2	18.3	21.9	16.3	25.3	17.3	23.9	16.3
Telugu	82.8	99.0	19.4	13.7	19.6	13.7	24.5	15.8	22.1	13.6
Thai	15.4	11.5	18.2	13.6	18.1	13.6	27.6	18.8	20.7	14.3
Turkish	10.4	8.4	28.6	17.3	28.7	17.5	31.2	16.1	30.9	16.9
Ukrainian	11.6	8.6	31.1	13.6	31.7	13.5	39.2	13.3	33.3	13.6
Urdu	28.2	22.6	42.3	22.7	33.1	20.1	46.4	25.1	36.9	20.5
Vietnamese	12.7	10.3	44.5	18.6	47.5	20.7	56.6	21.0	52.9	19.8
Welsh	40.8	33.0	48.9	20.8	49.0	20.8	54.9	21.4	51.4	20.9
Yoruba	105.1	94.8	61.2	49.7	61.9	49.4	62.7	50.2	64.2	49.4
	50.1	44.3	33.3	20.7	31.0	19.1	44.2	24.8	32.5	18.7

MMS Performance by Geographic Region (1107 Languages)

Region	Languages	Average CER	Languages with CER ≤ 5	Success Rate
Asia	335	1.6 ± 0.1	330	99%
South America	136	1.5 ± 0.2	132	97%
North America	144	2.2 ± 0.2	139	97%
Europe	41	1.7 ± 0.4	40	98%
Africa	363	2.9 ± 0.2	331	91%
Pacific	88	1.7 ± 0.5	87	99%

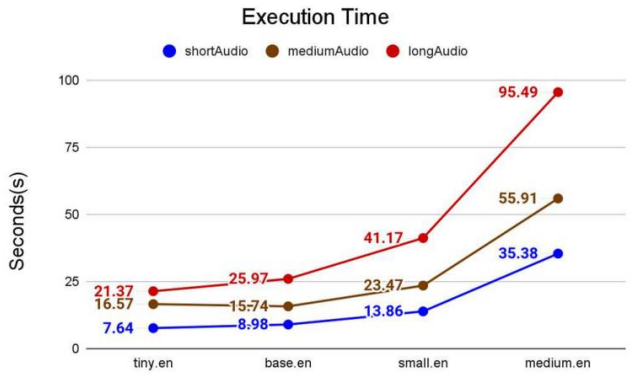
Total	**1,107**	**2.1 ± 0.1**	**1,059**	**96%**

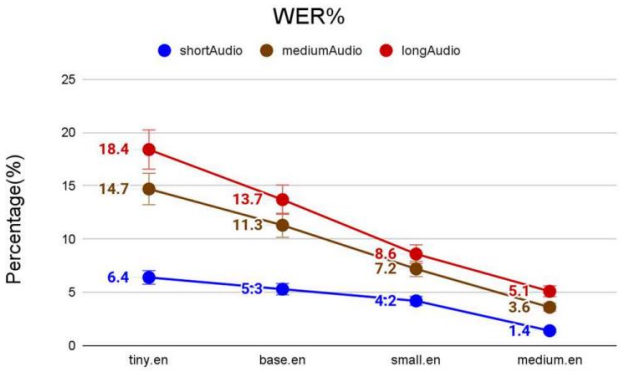
- Evaluation of the efficiency of state-of-the-art Speech Recognition engines

Comparison of open-source speech recognition tools

Model	English WER	French WER	English Inference Time	French Inference Time
DeepSearch	24.63%	29.53%	7.77s	2.86s
Kaldi	24.40%	-	3.97s	-
Vosk	18.84-26.82%	13.07%	1.27-3.34s	0.83s
LinTO	18,84%	11.77%	1.42s	0.94s

- A study of Audio-to-Text Conversion Software Using Whispers Model





The research suggests that the small.en model offers the best balance between accuracy and speed - it achieves WER% below 10% for all audio categories while requiring only about half the execution time of the medium.en model.

- WhisperSum: Unified Audio-to-Text Summarization

Experimental Results of the proposed WhisperSum model

Model	Precision	Recall	F1-Score
WhisperSum	91.98	92.12	92.04

- Speech Recognition Paradigms: A comparative Evaluation of SpeechBrain, Whisper and Wav2Vec2 Models

Comparison of Speech Recognition Models

Model	Avg. Similarity (%)	Avg. Jaccard Similarity (%)	Semantic Similarity (%)	Avg. WER (%)	Avg. CER (%)	BLEU Score
Speech Brain	95.57	99.36	97.33	5.43	6.58	79.32
Whisper	92.11	99.39	95.12	10.41	12.00	73.34
Wav2Vec2	84.51	99.49	90.19	29.01	32.91	66.45

- Evaluating Automatic Transcription Models Utilising Cloud Platforms

TABLE I
COMPARISON OF CLOUD TRANSCRIPTION SERVICES

ID	AssemblyAI	Speechmatics	OpenAI
1	0.10119	0.10714	0.05952
2	0.10119	0.09524	0.07738
3	0.10714	0.09524	0.07738
4	0.10714	0.08929	0.05952
5	0.09524	0.08333	0.07143
6	0.09524	0.09524	0.08929
7	0.08929	0.07143	0.05357
8	0.08929	0.08929	0.04762
9	0.07143	0.07738	0.04762
10	0.11905	0.10714	0.08929
11	0.08929	0.12500	0.04762
12	0.12500	0.10714	0.09524
13	0.09524	0.10119	0.05357
14	0.08929	0.07738	0.07738
15	0.10119	0.07738	0.07143
16	0.08333	0.08929	0.06548
17	0.10714	0.09524	0.05952
18	0.09524	0.09524	0.06548
19	0.13095	0.10119	0.08929
20	0.10119	0.08333	0.06548
21	0.16071	0.13690	0.08333
22	0.09524	0.08929	0.06548
23	0.07143	0.06548	0.04167
24	0.07143	0.06548	0.04167
25	0.09524	0.08929	0.04762
26	0.08929	0.05952	0.06548
27	0.10119	0.08333	0.06548
28	0.12500	0.08929	0.07143
29	0.11905	0.14286	0.09524
Avg	0.09208	0.08673	0.07246

This table shows the Word Error Rate (WER) for each of the 29 audio samples accross the three ASR services.

4. Conclusions

4.1 SLR process

Initially, research questions focused on comparing speech recognition models were defined aswell as relevant keywords for searching the articles.

Search strings captured relevant articles across databases including IEEE Xplore, Acm Digital Library, ScienceDirect and Scopus. The initial articles were reviewed in terms of having relevant title and abstract. Then articles were further reviewed using quality criteria and by looking whether the article contained relevant extractable data. The data was extracted and presented in the SLR.

4.2 SLR results

This SLR provided valuable insights into the current state of speech recognition technology, by comparing the performance of leading Speech Recognition systems.

It identified differences in accuracy and efficiency among leading ASR systems and established important relationships between size and transcription quality.

5. Literature

https://arxiv.org/abs/2305.13516 - Scaling speech technology to 1000+ languages

https://www.sciencedirect.com/science/article/pii/S1877050922014338 - Evaluation of the fficiency of state-of-the-art Speech Recognition engines

https://ieeexplore.ieee.org/document/10348186 - A study of Audio-to-Text Conversion Software Using Whispers Model

https://ieeexplore.ieee.org/document/10721926 - WhisperSum: Unified Audio-to-Text Summarization

https://ieeexplore.ieee.org/document/10544133 - Speech Recognition Paradigms: A comparative Evaluation Of SpeechBrain, Whisper and Wav2Vec2 Models

https://ieeexplore.ieee.org/document/10800465 - Evaluating Automatic Transcription Models Utilising Cloud Platforms

https://ieeexplore-1ieee-1org-10000076o0031.han.bg.pg.edu.pl/stamp/stamp.jsp?tp=&arnumber=10493971 - Automatic and Multilingual Speech Recognition and Translation by using Google Cloud API

https://arxiv.org/abs/2006.11477 - Wav2vec 2.0: A framework for Self-Supervised Learning of Speech Representations

https://ieeexplore.ieee.org/document/1318504 - Is word error rate a good indicator for spoken language understanding accuracy