SPEECH TO TEXT

CENL WEBINAR
AI IN LIBRARIES
20240926

Lars Flemming Mydtskov, it consultant
Lasse Rogers Nielsen, it developer
Ditte Laursen, senior researcher

DET KGL.
BIBLIOTEK
Royal Danish Library

MOTIVATION

CONSIDERATIONS

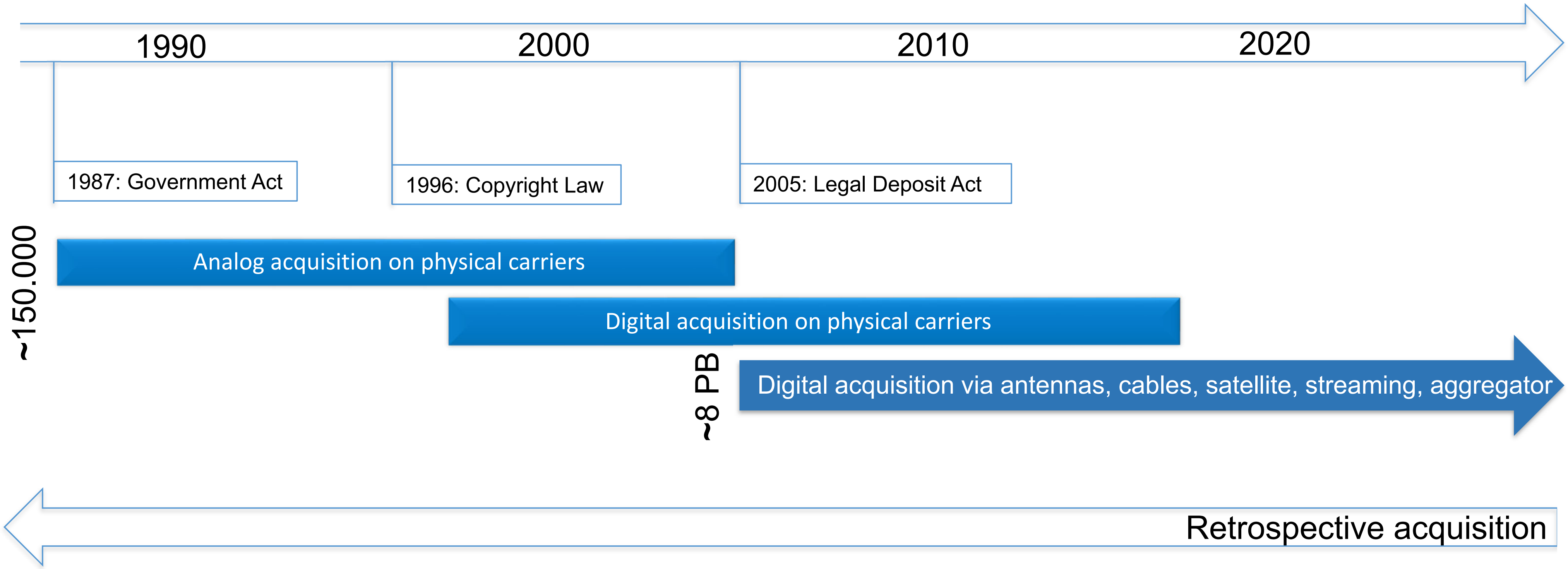
RESULTS

PERSPECTIVES

QA

DET KGL. BIBLIOTEK
Royal Danish Library

NATIONAL · REGIONAL · LOCAL RADIO/TV



1990

2000

2010

2020

1987: Government Act

1996: Copyright Law

2005: Legal Deposit Act

~150.000

Analog acquisition on physical carriers

Digital acquisition on physical carriers

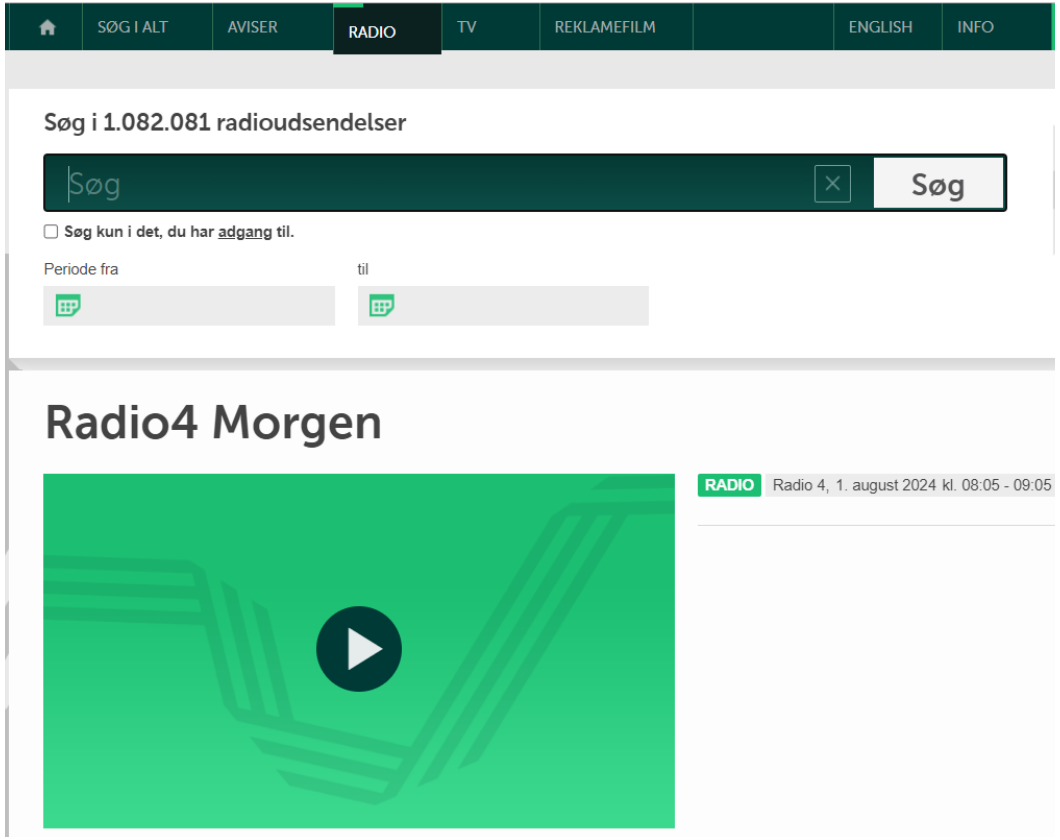
~8 PB

Digital acquisition via antennas, cables, satellite, streaming, aggregator

Retrospective acquisition

DET KGL.
BIBLIOTEK
Royal Danish Library

DIGITAL ACCESS FOR RESEARCH AND EDU



Søg i 1.082.081 radioudsendelser

Søg

□ Søg kun i det, du har adgang til.

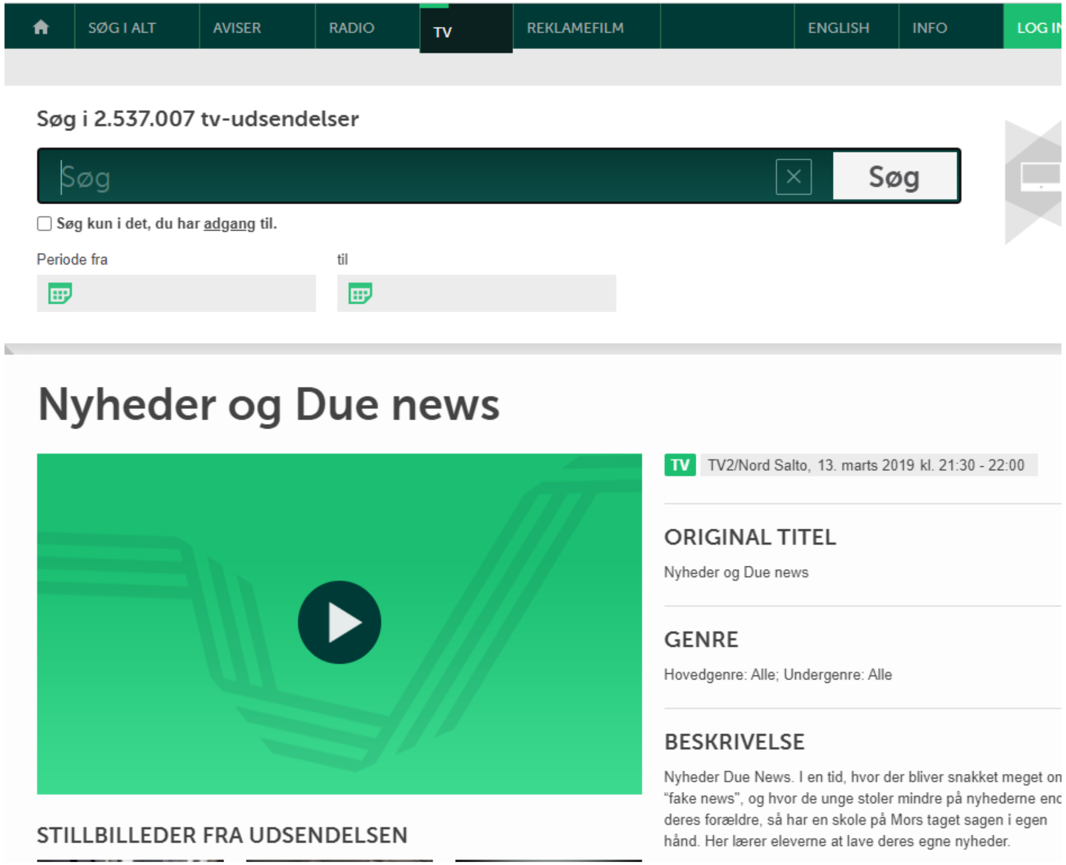
Periode fra

til

Radio4 Morgen

Radio 4, 1. august 2024 kl. 08:05 - 09:05

DIGITAL ACCESS FOR RESEARCH AND EDU



Søg i 2.537.007 tv-udsendelser

Søg | Søg

□ Søg kun i det, du har adgang til.

Periode fra

til

Nyheder og Due news

TV TV2/Nord Salto, 13. marts 2019 kl. 21:30 - 22:00

ORIGINAL TITEL

Nyheder og Due news

GENRE

Hovedgenre: Alle; Undergenre: Alle

BESKRIVELSE

Nyheder Due News. I en tid, hvor der bliver snakket meget om "fake news", og hvor de unge stoler mindre på nyhederne end deres forældre, så har en skole på Mors taget sagen i egen hånd. Her lærer eleverne at lave deres egne nyheder.

STILLBILLEDER FRA UDSENDELSEN

DET KGL. BIBLIOTEK
Royal Danish Library

DIGITAL ACCESS FOR RESEARCH AND EDU

Imagine the speech was searchable

Imagine it was possible to navigate to specific words or phrases within the audio content

SPEECH TO TEXT AS LEGAL DEPOSIT 2012-

—Public service content (ie. news)
—Danish produced content (ie. drama)
—Foreign content

Pre-produced and live-generated

00:00:32,059 –> 00:00:34,058 \n
Sådan lød det\n for under en time
siden.
00:00:34,359 –> 00:00:37,438 \n
USA’s præsident Trump\n aflyser et
historisk topmøde -
00:00:37,719 –> 00:00:40,558 \n -
med Nordkoreas leder, Kim Jong-un.

CONSIDERATIONS

DET KGL.
BIBLIOTEK
Royal Danish Library

COPYRIGHT

—Ownership of the original audio
—Reproduction rights

Text output
— Research okay; broader use would require a licensing agreement

GDPR

— Identifiable individuals
— Legal basis for processing
— Principle of data minimization
— Public interest exemption

Text output
— Research okay; broader use may spawn requests for access, correction, or deletion of transcribed data

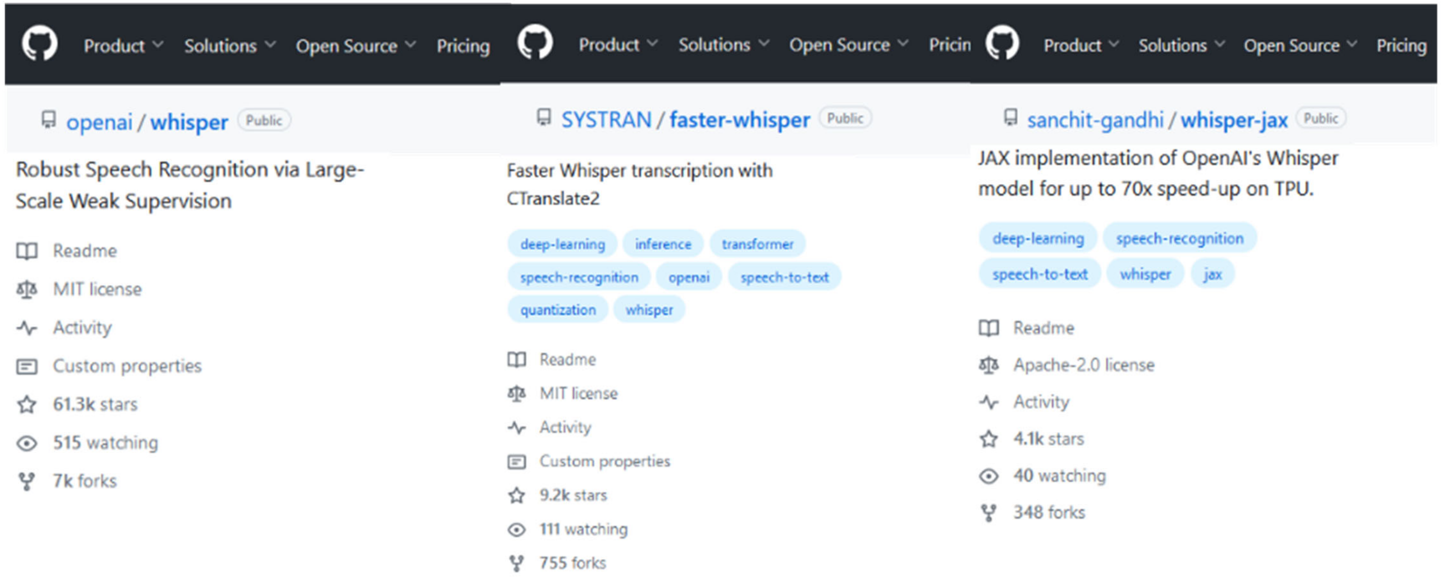
ETHICS

— Public interest vs. privacy rights
— Cultural evolution
— Accuracy and bias

RESULTS

DET KGL.
BIBLIOTEK
Royal Danish Library

WHIPSER IMPLEMENTATIONS



openai / whisper Public

Robust Speech Recognition via Large-Scale Weak Supervision

- Readme
- MIT license
- Activity
- Custom properties
- 61.3k stars
- 515 watching
- 7k forks

SYSTRAN / faster-whisper Public

Faster Whisper transcription with CTranslate2

deep-learning inference transformer speech-recognition openai speech-to-text quantization whisper

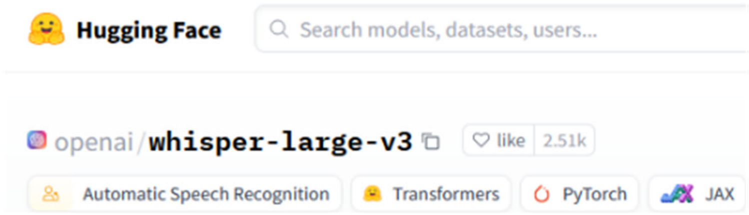
- Readme
- MIT license
- Activity
- Custom properties
- 9.2k stars
- 111 watching
- 755 forks

sanchit-gandhi / whisper-jax Public

JAX implementation of OpenAI's Whisper model for up to 70x speed-up on TPU.

deep-learning speech-recognition speech-to-text whisper jax

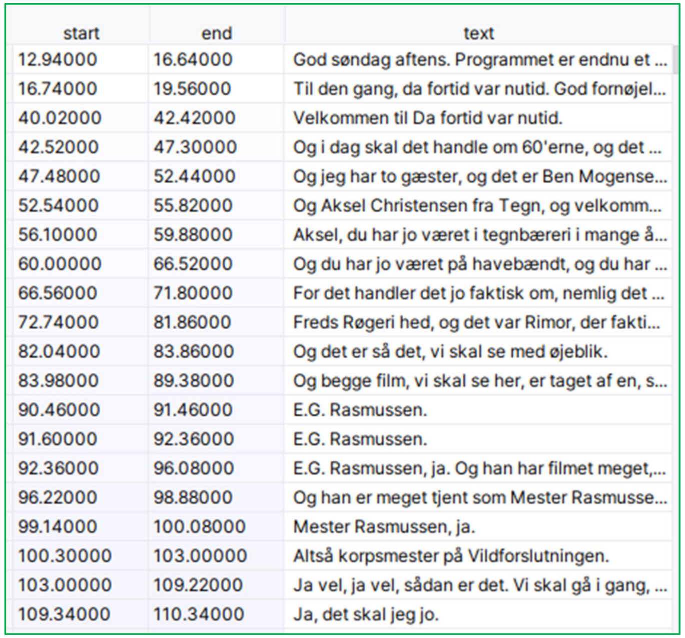
- Readme
- Apache-2.0 license
- Activity
- 4.1k stars
- 40 watching
- 348 forks



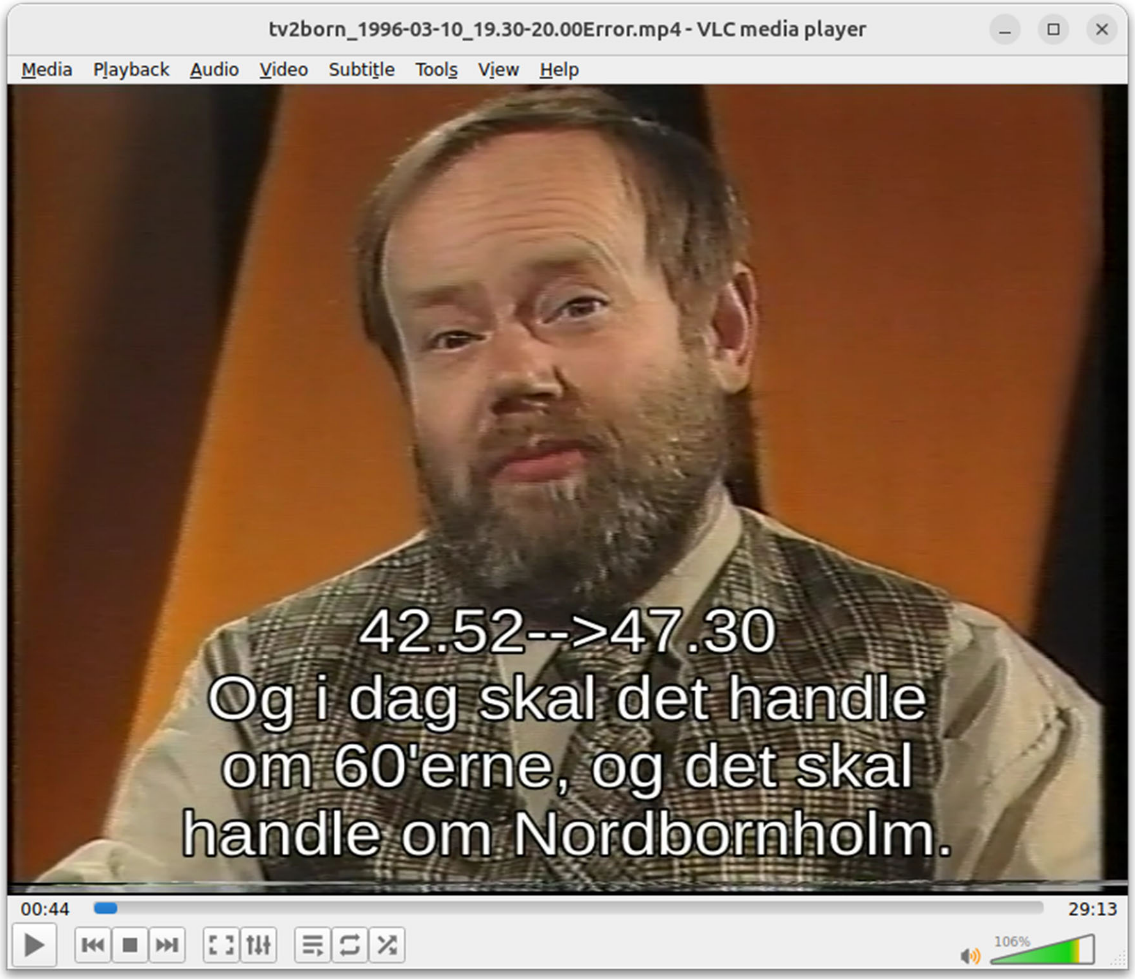
openai / whisper-large-v3 2.51k

Automatic Speech Recognition | Transformers | PyTorch | JAX

USE CASES - SUBTITLES



start	end	text
12.94000	16.64000	God søndag aftens. Programmet er endnu et ...
16.74000	19.56000	Til den gang, da fortid var nutid. God fornøjel...
40.02000	42.42000	Velkommen til Da fortid var nutid.
42.52000	47.30000	Og i dag skal det handle om 60'erne, og det ...
47.48000	52.44000	Og jeg har to gæster, og det er Ben Mogense...
52.54000	55.82000	Og Aksel Christensen fra Tegn, og velkomm...
56.10000	59.88000	Aksel, du har jo været i tegnbæreri i mange å...
60.00000	66.52000	Og du har jo været på havebændt, og du har ...
66.56000	71.80000	For det handler det jo faktisk om, nemlig det ...
72.74000	81.86000	Freds Røgeri hed, og det var Rimor, der fakti...
82.04000	83.86000	Og det er så det, vi skal se med øjeblik.
83.98000	89.38000	Og begge film, vi skal se her, er taget af en, s...
90.46000	91.46000	E.G. Rasmussen.
91.60000	92.36000	E.G. Rasmussen.
92.36000	96.08000	E.G. Rasmussen, ja. Og han har filmet meget,...
96.22000	98.88000	Og han er meget tjent som Mester Rasmusse...
99.14000	100.08000	Mester Rasmussen, ja.
100.30000	103.00000	Altså korpsmester på Vildforslutningen.
103.00000	109.22000	Ja vel, ja vel, sådan er det. Vi skal gå i gang, ...
109.34000	110.34000	Ja, det skal jeg jo.

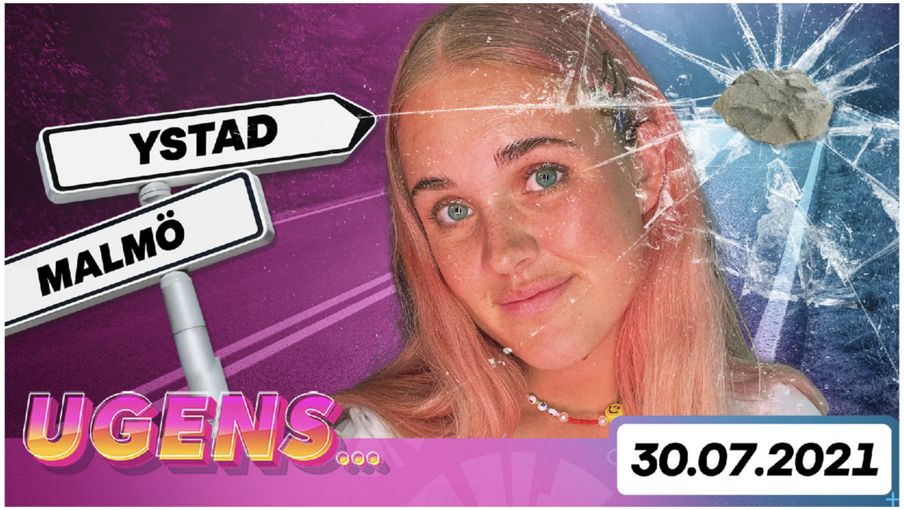


42.52-->47.30
Og i dag skal det handle
om 60'erne, og det skal
handle om Nordbornholm.

Digitized VHS video tape from 1996, subtitled for QA using whisper

USE CASES – TOPIC EXTRACTION

10 Translated keywords from TF-IDF algorithm performed on subtitles, displayed together with images from their corresponding program trailers



['bent', 'malmö', 'license plates', 'ystad', 'swedish', 'bornholm', 'police', 'stone', 'cars', 'sweden']

['good night', 'turkish', 'the boys', 'turkey', 'play', 'training', 'u16', 'mads', 'the hotel', 'the national team']

USE CASES – SPOKEN TIME

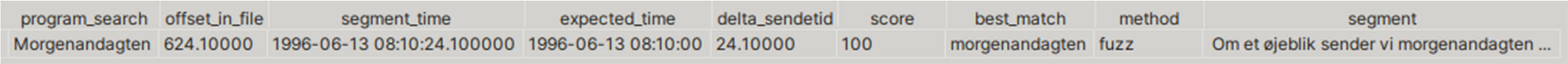
- What Time Is It?
  - search for "Klokken er" in transcript

```
01:54:18,46 --> 01:54:20,36
Og klokken er 11.
--
03:17:00,04 --> 03:17:00,76
Klokken er 12.20. Vi skal se på eftermiddagens
programmer.
--
03:58:12,5 --> 03:58:12,56
Klokken er 13.
--
06:01:57,82 --> 06:01:58,84
Klokken er 15.
--
07:04:03,16 --> 07:04:07,68
Ok, om 3 kvarter. Klokken er 16.
```

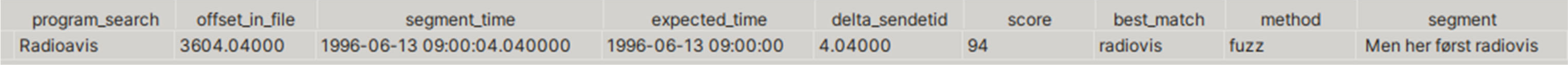
	Spoken Time	Offset in digitization (sec)
First spoken time	11:00 (11 am)	6.858
Last spoken time	16:00 (4 pm)	25.447
Elapsed seconds	18.000	18.589

USE CASES - SEGMENTATION

- Identifying start timestamps (search for announcement of expected program titles)



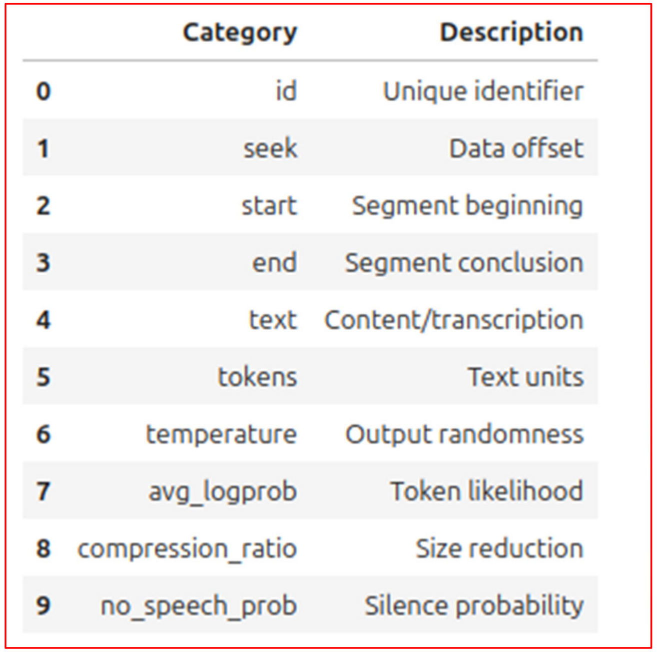
program_search	offset_in_file	segment_time	expected_time	delta_sendetid	score	best_match	method	segment
Morgenandagten	624.10000	1996-06-13 08:10:24.100000	1996-06-13 08:10:00	24.10000	100	morgenandagten	fuzz	Om et øjeblik sender vi morgenandagten ...



program_search	offset_in_file	segment_time	expected_time	delta_sendetid	score	best_match	method	segment
Radioavis	3604.04000	1996-06-13 09:00:04.040000	1996-06-13 09:00:00	4.04000	94	radiovis	fuzz	Men her først radiovis

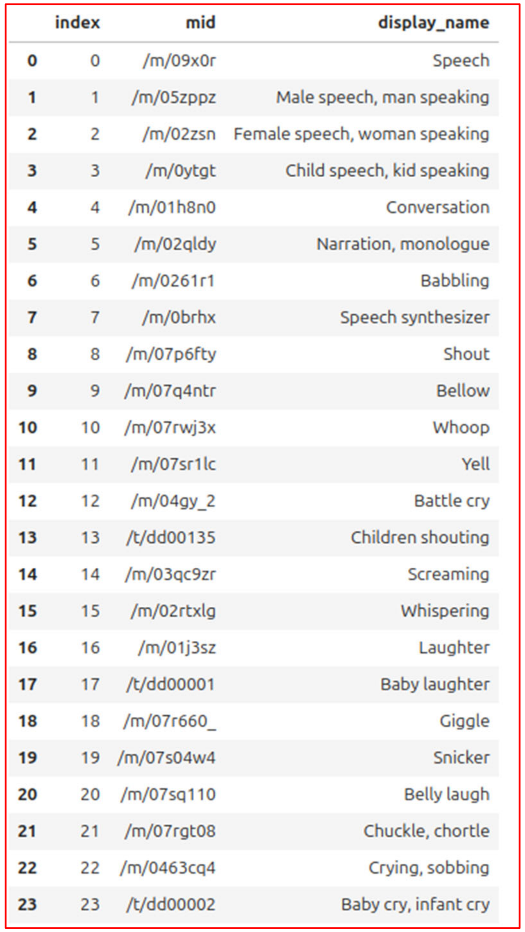
FEATURE ENGINEERING

Whisper Metadata



	Category	Description
0	id	Unique identifier
1	seek	Data offset
2	start	Segment beginning
3	end	Segment conclusion
4	text	Content/transcription
5	tokens	Text units
6	temperature	Output randomness
7	avg_logprob	Token likelihood
8	compression_ratio	Size reduction
9	no_speech_prob	Silence probability

YAMNet audio event classes



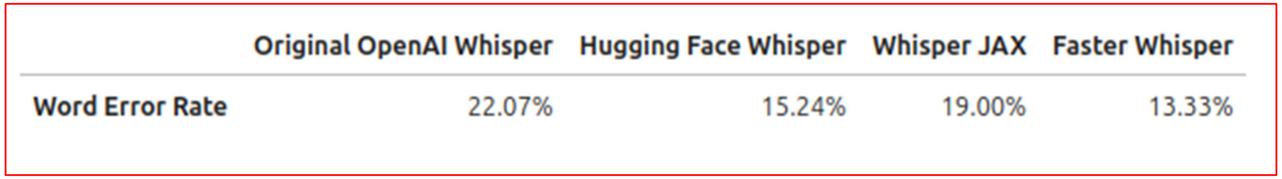
	index	mid	display_name
0	0	/m/09x0r	Speech
1	1	/m/05zppz	Male speech, man speaking
2	2	/m/02zsn	Female speech, woman speaking
3	3	/m/0ytgt	Child speech, kid speaking
4	4	/m/01h8n0	Conversation
5	5	/m/02qldy	Narration, monologue
6	6	/m/0261r1	Babbling
7	7	/m/0brhx	Speech synthesizer
8	8	/m/07p6fty	Shout
9	9	/m/07q4ntr	Bellow
10	10	/m/07rwj3x	Whoop
11	11	/m/07sr1lc	Yell
12	12	/m/04gy_2	Battle cry
13	13	/t/dd00135	Children shouting
14	14	/m/03qc9zr	Screaming
15	15	/m/02rtxlg	Whispering
16	16	/m/01j3sz	Laughter
17	17	/t/dd00001	Baby laughter
18	18	/m/07r660_	Giggle
19	19	/m/07s04w4	Snicker
20	20	/m/07sq110	Belly laugh
21	21	/m/07rgt08	Chuckle, chortle
22	22	/m/0463cq4	Crying, sobbing
23	23	/t/dd00002	Baby cry, infant cry

GROUND TRUTH CREATION

JiWER is a simple and fast python package to evaluate an automatic speech recognition system.

It supports the following measures:

- word error rate (WER)
- match error rate (MER)
- word information lost (WIL)
- word information preserved (WIP)
- character error rate (CER)

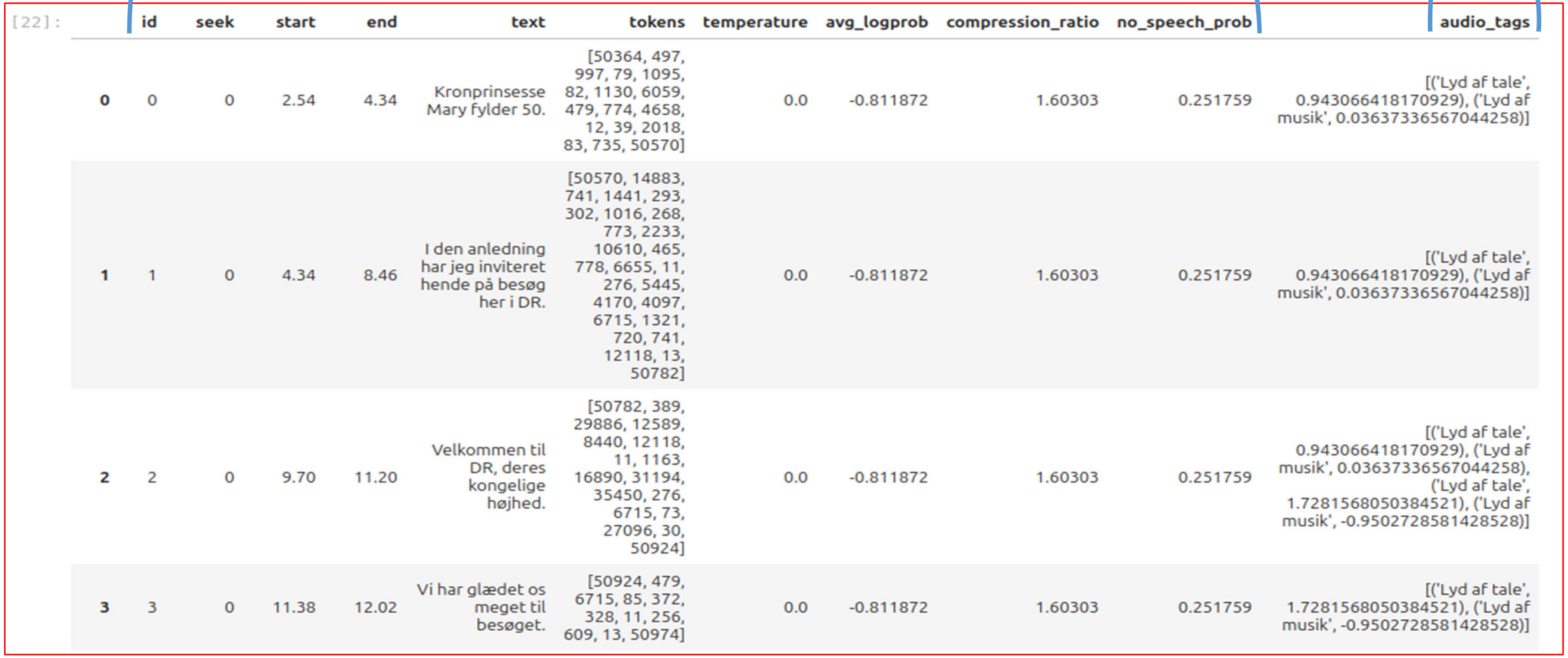


	Original OpenAI Whisper	Hugging Face Whisper	Whisper JAX	Faster Whisper
Word Error Rate	22.07%	15.24%	19.00%	13.33%

POSSIBLE TRAINING SETS

Whisper Metadata

YAMNet audio event classes



[22]:

	id	seek	start	end	text	tokens	temperature	avg_logprob	compression_ratio	no_speech_prob	audio_tags
0	0	0	2.54	4.34	Kronprinsesse Mary fylder 50.	[50364, 497, 997, 79, 1095, 82, 1130, 6059, 479, 774, 4658, 12, 39, 2018, 83, 735, 50570]	0.0	-0.811872	1.60303	0.251759	[('Lyd af tale', 0.943066418170929), ('Lyd af musik', 0.03637336567044258)]
1	1	0	4.34	8.46	I den anledning har jeg inviteret hende på besøg her i DR.	[50570, 14883, 741, 1441, 293, 302, 1016, 268, 773, 2233, 10610, 465, 778, 6655, 11, 276, 5445, 4170, 4097, 6715, 1321, 720, 741, 12118, 13, 50782]	0.0	-0.811872	1.60303	0.251759	[('Lyd af tale', 0.943066418170929), ('Lyd af musik', 0.03637336567044258)]
2	2	0	9.70	11.20	Velkommen til DR, deres kongelige højhed.	[50782, 389, 29886, 12589, 8440, 12118, 11, 1163, 16890, 31194, 35450, 276, 6715, 73, 27096, 30, 50924]	0.0	-0.811872	1.60303	0.251759	[('Lyd af tale', 0.943066418170929), ('Lyd af musik', 0.03637336567044258), ('Lyd af tale', 1.7281568050384521), ('Lyd af musik', -0.9502728581428528)]
3	3	0	11.38	12.02	Vi har glædet os meget til besøget.	[50924, 479, 6715, 85, 372, 328, 11, 256, 609, 13, 50974]	0.0	-0.811872	1.60303	0.251759	[('Lyd af tale', 1.7281568050384521), ('Lyd af musik', -0.9502728581428528)]

Future WER Labels

FINE TUNING

- Whisper can be fine-tuned locally to fit specific needs.
- The local fine-tuned layer serves as an addition to the public model
- You decide when, where and why you want to use a fine-tuned overlay (for dialects, time epochs, specific topic domains, etc.)

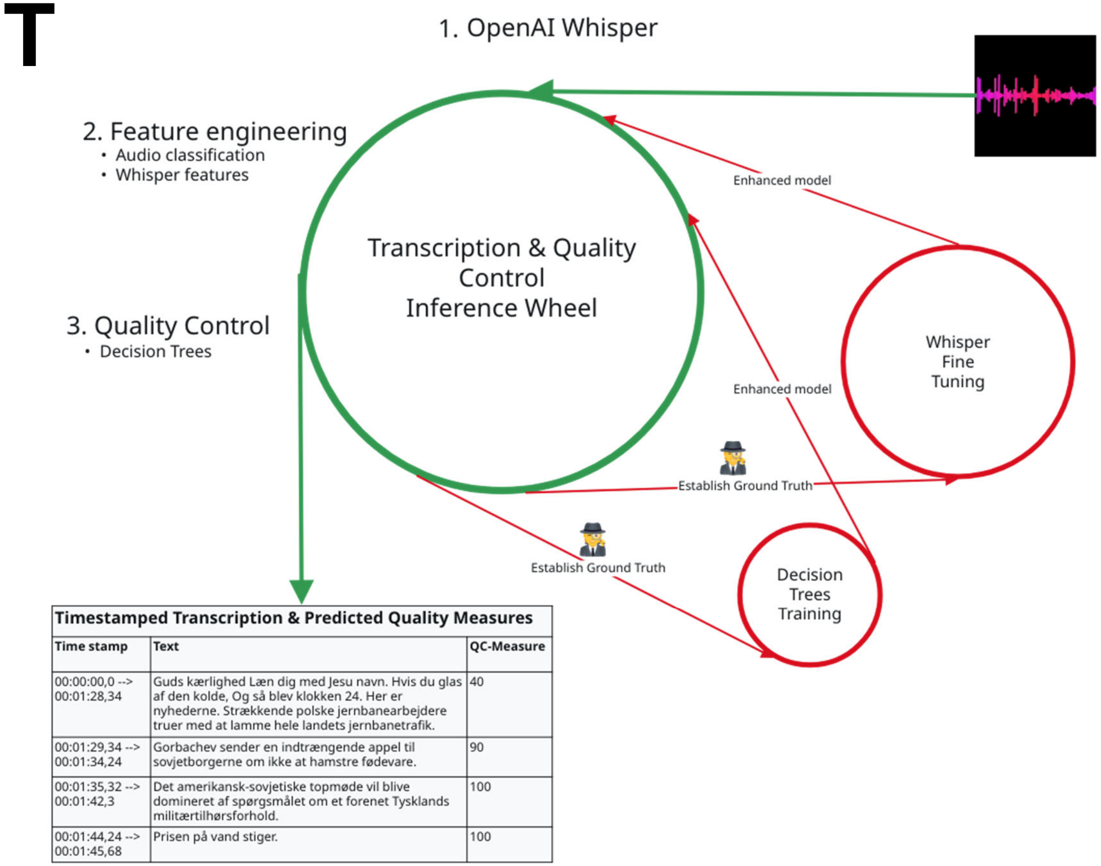
QUALITY ASSESMENT

Our experiments suggests that combining:

- Whisper Metadata & Audio Classification Classes with
- Ground Truth & Error measurement (WER, MER, WIL, WIP, CER)

allows for training of a Model aimed at Automated Quality Control, that can:

- predict quality of transcribed text segments,
- reveal the needs for fine tuning of Whisper and
- suggest which sound files to use for fine tuning.



Timestamped Transcription & Predicted Quality Measures

Time stamp	Text	QC-Measure
00:00:00,0 --> 00:01:28,34	Guds kærlighed Læn dig med Jesu navn. Hvis du glas af den kolde, Og så blev klokken 24. Her er nyhederne. Strækkende polske jernbanearbejdere truer med at lamme hele landets jernbanetrafik.	40
00:01:29,34 --> 00:01:34,24	Gorbachev sender en indtrængende appel til sovjetborgerne om ikke at hamstre fødevare.	90
00:01:35,32 --> 00:01:42,3	Det amerikansk-sovjetiske topmøde vil blive domineret af spørgsmålet om et forenet Tysklands militærtilhørsforhold.	100
00:01:44,24 --> 00:01:45,68	Prisen på vand stiger.	100

PERSPETIVES

DET KGL.
BIBLIOTEK
Royal Danish Library

PERSPECTIVES

Whisper part of a larger Toolbox

- Enrichment of metadata in large AV-collections
- Improve search and navigation i large AV-archives
- Identification of content
- Segmentation into programs
- Identification of replays

Easier access to more cultural heritage

MOTIVATION

CONSIDERATIONS

RESULTS

PERSPECTIVES

QA

DET KGL. BIBLIOTEK
Royal Danish Library

SPEECH TO TEXT

CENL WEBINAR
AI IN LIBRARIES
20240926

Lars Flemming Mydtskov, it consultant
Lasse Rogers Nielsen, it developer
Ditte Laursen, senior researcher

DET KGL.
BIBLIOTEK
Royal Danish Library