TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Oleksandra Zamana 211772IABM

# USING PRETRAINED LANGUAGE MODELS FOR IMPROVED SPEAKER IDENTIFICATION

Master's Thesis

Supervisor: Tanel Alumäe
PhD

Tallinn 2024

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Oleksandra Zamana 211772IABM

# EELKOOLITATUD KEELEMUDELITE KASUTAMINE KÕNELEJA TUVASTAMISE PARANDAMISEKS

Magistritöö

Juhendaja: Tanel Alumäe
PhD

Tallinn 2024

# Author’s Declaration of Originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Oleksandra Zamana

08.05.2024

# Abstract

This thesis explores improving named speaker identification using pretrained language models.

Aiming to correctly identify the speaker we use two approaches, with different types of models for each. In the supervised approach we use the textual information from each speaker's utterance in training data to finetune an encoder-based Roberta language model. As opposed to the supervised approach, when experimenting with large generative language models, such as GPT4 and GPT3, we perform zero-shot named speaker recognition using analogical text transcripts. In both supervised and zero-shot approaches, we perform experiments on two languages: the English VoxCeleb1 dataset and three Estonian broadcast news and conversational datasets. We compare results from textual data with those from audio data, which is a classic approach to solving this task. Last but not least, we interpolate text and audio-based models to establish if improvements can be made to state-of-the-art solutions.

Results of the work show that large language models are capable of improving named speaker identification performance dramatically, specifically when working with speech transcripts where speakers are introduced by their names. Moreover, there are instances where the OpenAI GPT-4 model outperforms human abilities in remembering the names of Estonian speakers mentioned in public debate transcripts.

The thesis is written in English and is 50 pages long, including 8 chapters, 16 figures, and 9 tables.

# Annotatsioon

## Eelkoolitatud keelemudelite kasutamine kõneleja tuvastamise parandamiseks

See lõputöö uurib nimelise kõneleja tuvastamise täiustamist eelkoolitatud keelemudelite abil.

Kõneleja õigeks tuvastamiseks kasutame kahte lähenemisviisi, millest igaühe jaoks on erinevat tüüpi mudeleid. Järelevalvega lähenemise korral kasutame iga kõneleja sõnavõttudest saadud tekstiteavet koolitusandmetes kodeerijapõhise Roberta keelemudeli peenhäälestamiseks. Erinevalt juhendatud lähenemisviisist teostame suurte generatiivsete keelemudelitega, nagu GPT4 ja GPT3, katsetamisel null-shot nimega kõnelejatuvastuse, kasutades analoogseid teksti transkripte. Juhendatud ja null-shot lähenemisviisides teeme katseid kahes keeles: inglise ja eesti keeles. Inglise keeles on kasutatud VoxCeleb1 andmestikku ning eesti keeles kolme eestikeelset ringhäälingu uudiste- ja vestlussaadete andmestikku. Võrdleme kontekstiandmete tulemusi heliandmetega, mis on selle ülesande lahendamise klassikaline lähenemine. Viimaks interpoleerime teksti- ja helipõhiseid mudeleid, et teha kindlaks, kas nüüdisaegseid lahendusi saab täiustada.

Töö tulemused näitavad, et suured keelemudelid on võimelised märkimisväärselt parandama nimetatud kõneleja tuvastamise jõudlust, eriti kui töötate kõne transkriptsioonidega, kus kõnelejaid tutvustatakse nende nimede järgi. Veelgi enam, on juhtumeid, kus OpenAI GPT-4 mudel ületab inimvõimed avaliku arutelu stenogrammides mainitud eesti keele kõnelejate nimede meelespidamisel.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 50 leheküljel, 8 peatükki, 16 joonist, 9 tabelit.

# List of Abbreviations and Terms

| | |
|---|---|
| DNN | Deep Neural Network |
| ASR | Automatic Speaker Recognition |
| GMM | Gausian Mixture Model |
| TDNN | Time Delay Neural Network |
| ROC | Receiver Operating Characteristic |
| MFCC | Mel Frequency Cepstrum Coefficient |
| LSTM | Long Short-Term Memory |
| LDE | Learnable Dictionary Encoding |
| BERT | Bidirectional Encoder Representations from Transformers |
| RoBERTa | Robustly Optimized BERT Approach |
| AA | Authorship Analysis |
| RNN | Recurring Neural Networks |
| LLM | Large Language Model |
| PCA | Principal Component Analysis |
| t-SNE | t-Distributed Stochastic Neighbor Embedding |
| GPT | Generative Pre-trained Transformer |
| POI | Person of Interest |

# Table of Contents

**1 Introduction** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **8**

**2 Related work** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **11**

2.1 Automatic Speech Recognition methods . . . . . . . . . . . . . . . . . 11

2.2 Content of speech and Automatic Speaker Identification . . . . . . . . . . 16

**3 Supervised speaker identification using text classification** . . . . . . . . . . **19**

3.1 Method . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

3.2 Experiments: VoxCeleb . . . . . . . . . . . . . . . . . . . . . . . . 20

3.2.1 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20

3.2.2 Experiments . . . . . . . . . . . . . . . . . . . . . . . . . . . 22

3.2.3 Results and analysis . . . . . . . . . . . . . . . . . . . . . . . 23

3.3 Experiments: Estonian broadcast and public debate speech . . . . . . . . 23

3.3.1 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23

3.3.2 Experiments . . . . . . . . . . . . . . . . . . . . . . . . . . . 27

3.3.3 Results and analysis . . . . . . . . . . . . . . . . . . . . . . . 28

**4 Zero-shot speaker identification with LLMs** . . . . . . . . . . . . . . . . . **30**

4.1 Method . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30

4.2 Experiments: VoxCeleb . . . . . . . . . . . . . . . . . . . . . . . . 30

4.2.1 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30

4.2.2 Experiments . . . . . . . . . . . . . . . . . . . . . . . . . . . 33

4.2.3 Results and analysis . . . . . . . . . . . . . . . . . . . . . . . 33

4.3 Experiments: Estonian broadcast and public debate speech . . . . . . . . 34

4.3.1 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

4.3.2 Experiments . . . . . . . . . . . . . . . . . . . . . . . . . . . 35

4.3.3 Results and analysis . . . . . . . . . . . . . . . . . . . . . . . 36

**5 Summary** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **42**

**References** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **43**

**Appendix 1 – Non-Exclusive License for Reproduction and Publication of a Graduation Thesis** . . . . . . . . . . . . . . . . . . . . . . . . . . . **48**

**Appendix 2 – Example of Full News Transcript from Estonian dataset** . . . . . **49**

# List of Figures

1 Deep learning for speaker recognition components. . . . . . . . . . . . . 11
2 Transformer architecture. . . . . . . . . . . . . . . . . . . . . . . . . . 14
3 BERT architecture. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
4 RoBERTa architecture. . . . . . . . . . . . . . . . . . . . . . . . . . . 15
5 Example of a few-shot learning. . . . . . . . . . . . . . . . . . . . . . . 17
6 Example of a zero-shot learning. . . . . . . . . . . . . . . . . . . . . . . 17

7 VoxCeleb automated data processing pipeline. . . . . . . . . . . . . . . . 20
8 ArvamusFestival session. . . . . . . . . . . . . . . . . . . . . . . . . . 25
9 Visualization of Estonian Speaker Embeddings: Clustering of Speakers Based on Speech Characteristics. . . . . . . . . . . . . . . . . . . . . . . . 26
10 Interpolation of the RoBERTa and audio models. . . . . . . . . . . . . . 28

11 Length of utterances of the transcribed VoxCeleb1 dataset. . . . . . . . . 32
12 Length of utterances of the transcribed Estonian dataset. . . . . . . . . . . 35
13 Interpolation of GPT4 and audio models. . . . . . . . . . . . . . . . . . . 36
14 Example of outstanding GPT-4 performance. . . . . . . . . . . . . . . . . 39
15 Outline of LLM-based speaker identification of VoxCeleb1 dataset. The instruction prompt is slightly shortened. . . . . . . . . . . . . . . . . . . 40
16 Outline of LLM-based speaker identification of broadcast news and multi-party conversations. The instruction prompt is slightly shortened. . . . . . 41

# List of Tables

1 Goal of the work: Speaker Identification employing content of speech. . . 9

2 Results on VoxCeleb1 dataset with supervised models. . . . . . . . . . . 23
3 Training, development and test data for Estonian experiments. For each source, the number of recordings and the total amount in hours is given. . 23
4 Speaker identification precision (P) and recall (R) rates of different models on Estonian test sets. . . . . . . . . . . . . . . . . . . . . . . . . . . 29

5 Example predictions of the GPT-4 model with true speakers. . . . . . . . 31
6 Results on VoxCeleb data with GPT models. . . . . . . . . . . . . . . . . . 34
7 Results on VoxCeleb data with GPT models, when provided the 10 most probable speakers from the audio-based model. . . . . . . . . . . . . . . 34
8 Precision and recall rates of LLM-based speaker identification on the Estonian test sets. We compared OpenAI's GPT3.5 (with 16k token context size) and GPT4 (with 128k context size). . . . . . . . . . . . . . . . . . 37
9 Speaker identification precision and recall on Estonian opinion festival transcripts with GPT-4, based on ASR transcripts *vs* reference transcripts, and with increasing name comparison edit distance. . . . . . . . . . . . . 38

# 1. Introduction

The problem of speaker recognition has been researched since at least as far back as the 1960s [1]. As a speaker's voice inherently reflects personal traits due to the unique characteristics of their pronunciation organs and speaking manner, such as the individual vocal tract shape, larynx size, accent, and rhythm [2], it makes it possible to automatically identify speakers based on their speech. This process is called automatic speaker recognition (ASR). Speaker recognition has diverse applications in various practical settings. Examples include authentication for personal gadgets, security of transactions for bank trading and remote payments, forensics for establishing guilt or innocence, surveillance, and finally audio-based information retrieval for broadcast news, meetings, and calls [3]. The main possible use-case for this work is indexing large audio archives, like ones of the ERR, radio companies, or companies that have large amounts of saved meeting recordings. The results of this work would allow speaker-based retrieval from such archives, as well as improved transcription and labeling systems. Speaker-based retrieval is a process of searching and retrieving audio or video content from an archive based on the identity or characteristics of the speaker(s) in the content.

The problem that this thesis is trying to solve lies in the dismissal of the content of speech in the state-of-the-art speaker identification models. In modern works, speaker identification is performed majorly by leveraging auditory cues from the speaker's voice. Such systems are preferred since they yield high accuracy, especially after the improvements of recent years. Indeed, voice is a rich source of unique information and can provide a strong basis for identification: each of us features distinct spectral details, intonation, and stress, as well as tempo, volume, and rhythm of speech. Converting this information into speaker embeddings produces high-dimensional numerical vectors that capture these various characteristics of the input audio. The uniqueness of the properties makes embeddings from different speakers easily distinguishable.

Despite the obvious advantages of the audio-based approach, there are cases when paying attention to the content of the speech might yield an increase in the identification accuracy. When recognizing the speaker in everyday life, people consider the acoustic details and the content of speech. We pick up on contextual clues from background information about the speaker, such as topics they are likely to engage in and their choice of words or sentence structure. Language can convey much more than simply a semantic meaning and it can in many cases significantly improve a discriminative portfolio of a speaker. We furthermore observe the importance of the content in the scenarios where we have limited or scarce background information about a speaker, in those cases, identification almost solely relies on the possible initial introduction of the speaker, either by the speaker or by a third party. From these examples, we can conclude that the content of speech should not be dismissed and can become crucial in the case of both abundant initial information about the speaker and when the deduction relies solely on the introduction.

This work was motivated by the possibility of improving users' digital portfolios, as it allows people to receive more accurate and personalized services in various fields. This can improve accessibility and quality of services in both the physical and cyber worlds.

Table 1. Goal of the work: Speaker Identification employing content of speech.

| True Speaker | Speech Segment |
|---|---|
| Uku Toom | Mina olen Uku Toom. Valitsus arutab ÜRO ränderaamistiku teemat homsel kabineti nõupidamisel ja teeb kindlasti ka otsuse, ütles infotunnis peaminister Jüri Ratas... |
| Madis Hindre | Maksu- ja Tolliameti kodulehel seisab teade, mis ütleb, et tavaremondi tegemisel ei ole õigust eluasemelaenu intresse tulust maha arvata. Õiguskantsleri nõuniku Ago Pelisaare sõnul ei vasta see tõele... |
| Kersti Kaljulaid | Plaan on siis see, et aasta lõpuks on meil e-residentsus kaks punkt null plaan koos isegi natuke varem, detsembri alguseks ja see tuleb läbi juhtrühmade, kes siis tegelevadki nende valdkondadega näiteks e-residentsus... |
| Uku Toom | Vabariigi president Kersti Kaljulaid nimetas ametisse Eesti Panga uue presidendi, selleks on panga endine asepresident Madis Müller. Tõnu Karjatse käis Kadriorus... |
| Kersti Kaljulaid | No tegelikult ei ole pangad ja pangaliit ja pangandussektori inimesed öelnud, et e-residentsus on spetsiifiline probleem, me teame, et rahapesu kui selline on probleem... |
| Kaja Kallas | Õigusriigis tuleb käituda seaduse järgi, see on väga oluline majanduskeskkonnale. Kui on mingid reeglid kokku lepitud, siis poolel teel neid reegleid kuidagi muuta ei saa... |

The goal of the work, as illustrated in Table 1, is to explore practically if content improves the accuracy of audio-based speaker identification while leveraging pretrained language models, as well as Large Language Models. We base this work on real-life interactions, such as interviews, talk shows, or debate speeches, and employ pre-trained models that are then finetuned to assess the results. To have the opportunity for comparison and more diversity, a total of four datasets are utilized in this work: three Estonian datasets, that are merged to create one, and a single Voxceleb1 English dataset. Chronologically in this work we first worked with English datasets and then proceeded to repeat experiments on Estonian datasets. In the early stages of the work, we extracted speech transcripts from the initial audio datasets and used them to train a purely text-based Roberta model. After this, we extracted speaker embeddings from the same datasets to train an audio-based model. After estimating the results independently we proceeded to integrate the results from two models to see if we discovered that the content-based model further amplifies the audio-based model.

Then we further explore how Large Language Models (LLMs), such as GPT models, can be used to identify the speakers when given transcribed utterances from the speech. Models are not trained for this specific task, which makes it a zero-shot identification task, and we do not restrict the pool of possible candidates, which means it is also an open-set identification task. Model is normally able to identify individuals with notable online presence and accuracy is limited. Indeed, we show how impressively LLMs handle the speaker identification tasks, especially when provided a full transcript with introductions included, as seen in broadcast news, talk shows, and panel discussions. On the test set of Estonian radio talk shows, GPT-4 improves the recall rate of speaker identification from 52% of the audio-based model to 98%, while achieving 100% precision.

# 2. Related work

## 2.1 Automatic Speech Recognition methods

Deep Neural Networks (DNNs) have recently become a state-of-model solution for several tasks, such as speaker identification, verification, diarisation, emotion recognition, and so on, boasting greater performance compared to the previous approaches. Before DNNs entered a modern speech processing scene for a long time Gaussian mixture models (GMMs) and i-vector approaches [4] have been employed to solve these tasks. GMM-based speaker recognition frameworks first statistically construct the acoustic features extracted from speech signals and then model the probability distribution of the features using a mixture of Gaussian distributions. Factor analysis-based Eigenvoice (i-vector) strategies use i-vectors to represent speaker features and compare the distances of similarities between these vectors to identify speakers. As shown in both [4] and [5] i-vector-based models perform better than GMM-based models, especially when tested in challenging environments (background noise, channel mismatch, etc).

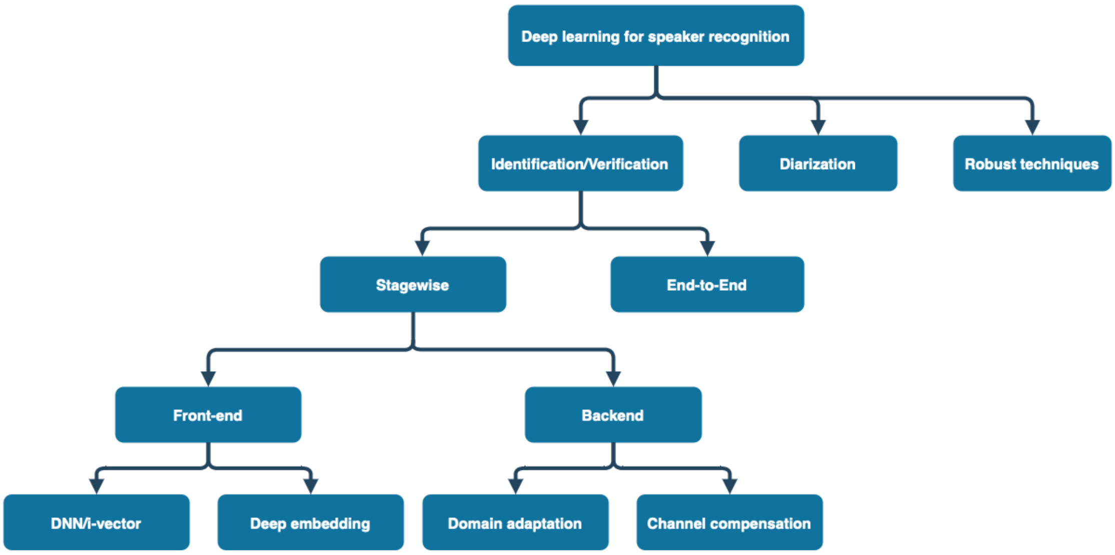


Figure 1. Deep learning for speaker recognition components.

Figure 1 based on [3] gives an overview of what kind of components of the speaker recognition process can influence the results when working with DNNs. Speaker recognition is divided into speaker identification and speaker verification. Speaker verification verifies if a claimed identity matches the actual identity, while speaker identification determines the identity of an unknown speaker from a set of known or potential identities. Verification involves a binary decision process, while identification requires selecting the most similar reference template from multiple candidates. Both speaker verification and identification algorithms can be divided into two main groups: Stagewise and End-to-End. [3]

In the case of the Stagewise algorithms, there is usually a front-end for the extraction of speaker features and a back-end for the similarity calculation of speaker features. The front-end aims to learn speaker representations, there are two main approaches: DNN/i-vector or a deep embedding. They differ in their approaches to feature extraction, representation learning, and model architecture. DNN/i-vector front end combines deep learning and statistical modeling techniques and relies on phonetic labels, while the deep embedding front end relies exclusively on deep learning methods for speaker representation learning and does not rely on phonetic labels. In the case of the deep embedding front-end, we can choose different networks and structures, temporal pooling layers, and objective functions. [3]

Unlike stage-wise techniques, end-to-end speaker verification takes a pair of speech utterances as the input and produces their similarity score directly. The main difference between end-to-end speaker verification and the deep embedding techniques in stage-wise speaker verification is the loss function. [3]

In recent years studies started introducing DNNs into the classic approach, for example in [6] DNN was used to enhance traditional i-vectors. Later DNNs start playing the main role in the scientific works, for example for speaker verification [7] and for robust speaker recognition [8]. Efforts have been made to further improve deep embedding learning. The loss function is used to discriminate between speakers, and examples of its improvements can be seen in [9], where discriminant analysis loss is designed for end-to-end training and improved the state-of-art results significantly, and in [10], where softmax-based cross-entropy loss function with adaptive parameters (ParAda) also improves the accuracy. In another paper [11] authors are focusing on improving the loss function for the speaker verification task, coming up with the maximization of the partial area under the Receiver-operating-characteristic (ROC) curve (pAUC) for deep embedding-based text-independent speaker verification. Examples of improvements to the aggregation functions are the works [9] and [12].

The combinations of network input representations and network structures are diverse. Some examples are [13], which uses acoustic features equipped with the techniques of spectral centroids, group delay function, and integrated noise suppression. [14] is using raw waves and CNN SincNet architecture. [15] is introducing factorized TDNN (F-TDNN) which gives substantial improvements over TDNNs. [12, 16] are examples of using spectrograms as an input along with ResNet architecture. In [17] Wav2Spk architecture is created, where speaker embeddings are derived from waveforms using a feature encoder, and a temporal gating unit along with an instance normalization scheme are used. In [11] MFCC embeddings and TDNN are used. Finally, in [18] raw waves and CNN-LSTM architecture are used.

The temporal pooling layer is a bridge between the frame-level and utterance-level hidden layers [17]. As an example in [12] the authors applied a dictionary-based NetVLAD and GhostVLAD layer to aggregate features across time. In another research [19] a learnable dictionary encoding (LDE) pooling layer is used. LDE was proposed in the research [20].

Pretrained language models are considered to be the latest and greatest in the speech-processing world. Pretrained language models are deep learning models trained on vast amounts of text data. Examples include OpenAI's GPT (Generative Pretrained Transformer) series, Google's BERT (Bidirectional Encoder Representations from Transformers), and Facebook's RoBERTa (Robustly optimized BERT approach). These models have been pretrained on large-scale text corpora and can be further fine-tuned for various natural languages processing tasks such as text generation, translation, sentiment analysis, and more.

Even though these models are relatively new, researchers already turned their attention to them. One of the most transformative breakthroughs was developing a transformer encoder model [21]. This model is based solely on attention mechanisms, unlike the earlier state-of-art models that relied on recurrence and convolutions and involved both encoders and decoded, often connected via attention mechanism. Transformer is a fundamental part of OpenAI GPT architecture, along with unsupervised pre-training [22]. GPT models are decoder-only models that generate text. The first stage of the GPT system is to train the model on a very large amount of data in an unsupervised manner (which means that training data does not have any specific task labels or annotations). In the second stage the model is fine-tuned on much smaller supervised datasets to help it solve specific tasks [22][23].

Figure 2 from [21] depicts classic transformer architecture. Architecture is based on a multi-layered structure of encoders and decoders that utilize self-attention mechanisms to process sequences of data. Each encoder layer consists of self-attention and feed-forward neural networks, and each decoder layer includes an additional attention mechanism that focuses on the encoder's output. This design allows the model to handle complex sequence-to-sequence tasks by dynamically weighing the importance of different parts of the input data, enabling effective handling of long-range dependencies without relying on recurrent networks.

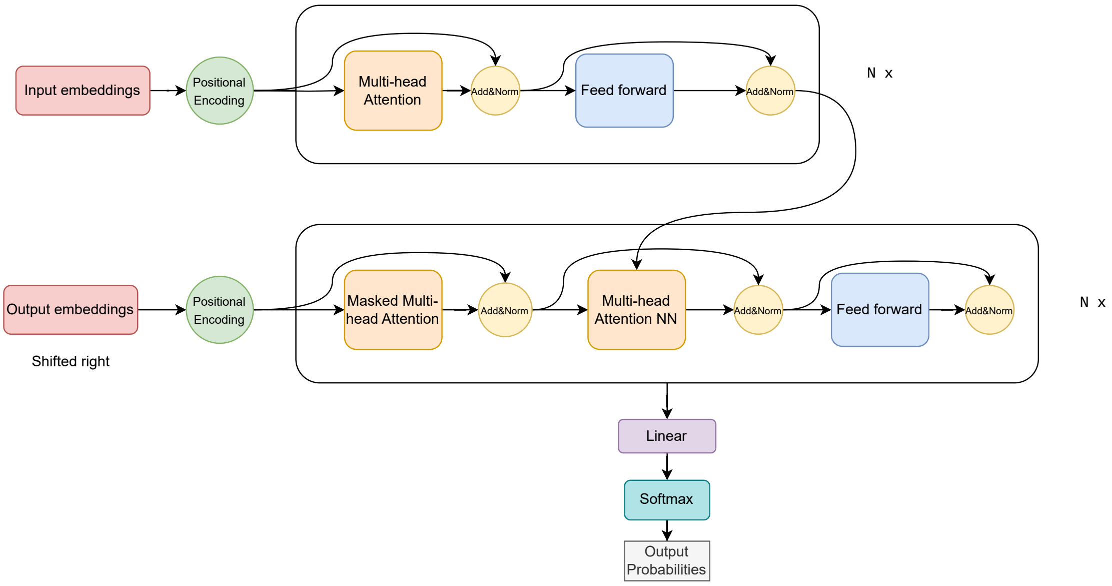


Figure 2. Transformer architecture.

Conversely, the BERT model originated from improvements that pretraining bidirectional contextual representations introduced [24], and it is designed to pretrain deep bidirectional representations from the unlabeled text by joint conditioning on both left and right context in all layers, unlike in the unidirectional left-to-right model pertaining. As depicted in Figure 3, BERT leverages the encoder component of the Transformer architecture, enhanced by a self-attention mechanism that enables bidirectional context processing, crucial for understanding the full context of a sentence. It is pre-trained using two specific tasks: Masked Language Modeling (MLM), where random tokens are masked and predicted by the model, and Next Sentence Prediction (NSP), where the model guesses if one sentence logically follows another. BERT's input representation combines token embeddings, which capture the semantic meaning of each word, with positional embeddings, which encode the position of each token in the sequence, and segment embeddings, which differentiate between sentences in tasks involving pairs of sentences. This combination allows BERT to maintain sequence order and distinguish between different segments effectively. The attention mechanism in BERT processes each token in the context of all other tokens in the sequence, making the model inherently bidirectional and capable of generating rich, context-aware embeddings. BERT was originally implemented in two model sizes: bert-base: 12 encoders with 12 bidirectional self-attention heads totaling 110 million parameters, and bert-large: 24 encoders with 16 bidirectional self-attention heads totaling 340 million parameters. Both models were pre-trained on the Toronto BookCorpus (800M words) and English Wikipedia (2,500M words).

Finally, an optimized version of the BERT model was proposed, RoBERTa (Robustly Optimized BERT Approach) [25]. The model was created as a result of key hyperpa-

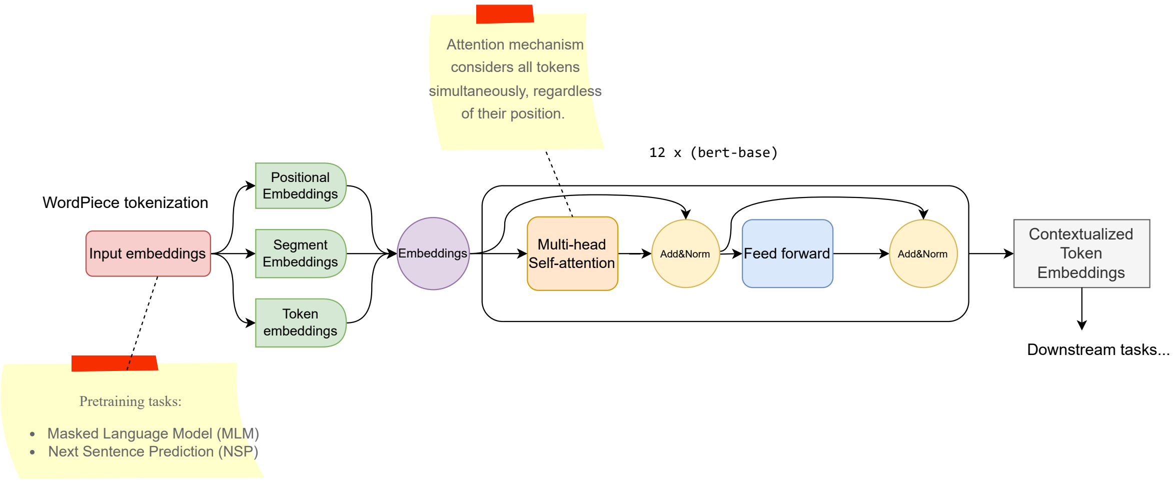

Attention mechanism considers all tokens simultaneously, regardless of their position.
12 x (bert-base)
WordPiece tokenization
Input embeddings
Positional Embeddings
Segment Embeddings
Token embeddings
Embeddings
Multi-head Self-attention
Add&Norm
Feed forward
Add&Norm
Contextualized Token Embeddings
Downstream tasks...
Pretraining tasks:
Masked Language Model (MLM)
Next Sentence Prediction (NSP)

Figure 3. BERT architecture.

rameters and training data size analysis and surpassed the published results of BERT. RoBERTa builds upon the BERT architecture but introduces several key optimizations to enhance performance. Unlike BERT, RoBERTa removes the Next Sentence Prediction (NSP) pre-training task, focusing solely on the Masked Language Model (MLM) task with dynamically changed masking patterns, rather than the static masking used in BERT. It is trained on a much larger and more diverse dataset, and with significantly more data and longer training times, which helps improve the model's language understanding capabilities. RoBERTa also utilizes larger batch sizes and longer sequences during training and fine-tunes the hyperparameters more extensively than BERT. These modifications enable RoBERTa to achieve better performance and more robustness across a wider range of natural language processing tasks compared to its predecessor. Figure 4 describes RoBERTa architecture and visualizes similarities between BERT and RoBERTa.

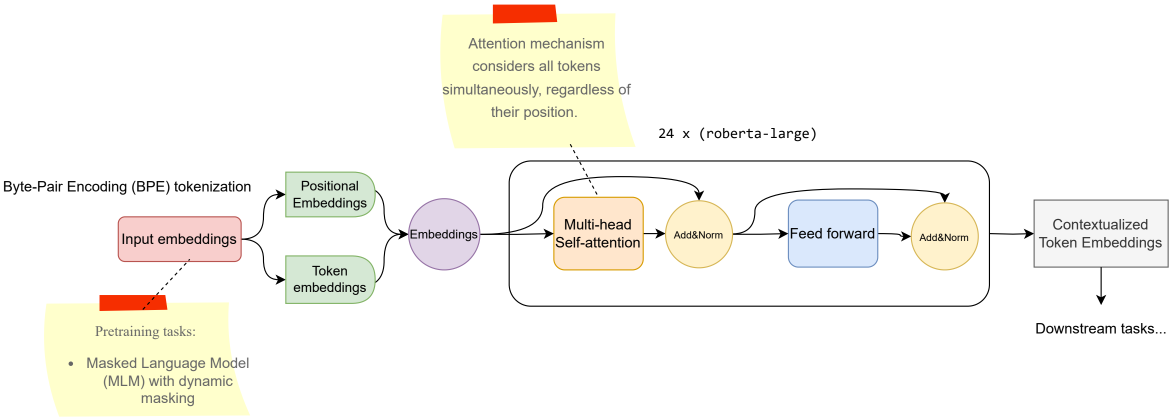

Attention mechanism considers all tokens simultaneously, regardless of their position.
24 x (roberta-large)
Byte-Pair Encoding (BPE) tokenization
Input embeddings
Positional Embeddings
Token embeddings
Embeddings
Multi-head Self-attention
Add&Norm
Feed forward
Add&Norm
Contextualized Token Embeddings
Downstream tasks...
Pretraining tasks:
Masked Language Model (MLM) with dynamic masking

Figure 4. RoBERTa architecture.

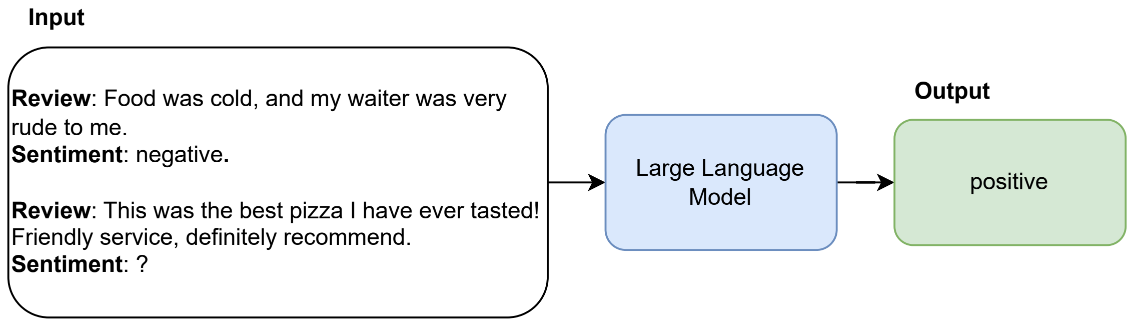
## 2.2 Content of speech and Automatic Speaker Identification

Content of speech has received limited attention in the ASR in the past works. Some research has tried building various linguistic-based systems to identify named speakers with the ultimate goal of speaker diarization (associating speaker identities with homogeneous audio segments) in French news. In work [26] authors use the most common patterns in which the names of the current (who is speaking), previous (who just spoke), or next (who will speak) speakers occur in the news segments to identify speakers. This is an example of manually built rules. The next generation of research, with works such as [27, 28, 29], uses semantic classification trees to identify speakers. Semantic classification trees (SCT) are hierarchical structures based on regular expressions that organize words or concepts based on their semantic relationships and similarities. The goal in these works, analogically, was to categorize full name pairs into four categories (previous, current, next, and other). In studies [30, 31] SCT and acoustic systems are combined with belief functions (mathematical frameworks that represent uncertainty and assign degrees of belief to propositions or hypotheses) to achieve diarization. [32] unifies speaker diarization and speaker clustering steps and creates a "person instance graph" framework. In [33] authors propose a conditional maximum entropy framework that combines linguistic and acoustic features. On the other hand, in [34] probabilistic latent semantic indexing (PLSI) is used to model a speaker's vocabulary, which is believed to reflect their identity. Based on such vocabulary an identity score is created that is later combined with the acoustic-based score.

Another domain where the importance of content can not be overlooked is authorship verification and attribution, also known as Authorship Analysis (AA), [35]. In much the same way as broadcast news speakers have vocabularies that closely reflect their identity, each writer has a unique writing style. They might have very specific speech patterns, use different word frequencies, etc., which distinguish their penmanship from others and can be used to predict the authorship of unidentified documents based on some other documents. A multitude of methods have been proposed to solve this task, including Stylometric Analysis (static stylometrics, dynamic n-grams), topic modeling and semantic analysis, and various word embeddings (distributed bag of words version of paragraph vector, word2vec representations) [36]. In more recent years researchers started making use of Machine Learning technologies, such as support vector machines (SVMs) and random forests [37] to learn the sociolinguistic characteristics and map them to the authors. Of course, the latest method for authorship analysis is deep neural networks. In [38], recurring neural networks (RNNs) were used for language modeling and subsequent document classification while in [39] a modification of a BERT model called BertAA was proposed.

In speech emotions are represented both via the vocal features (pitch, tone, etc.) and via the words that are associated with certain feelings, this research focusing on the task of speech emotion recognition often combine audio and text features. Examples of research based solely on acoustic features exist [40], but more frequently context is used to improve the accuracy. In [41] LSTM was used to extract acoustic features and a convolutional model was used to extract information from word sequences, while in [42] acoustic ASR is fine-tuned on emotion-annotated speech data. Motivated by promising results of the self-supervised learning methods (SSLs) many authors tried applying them to the problem of emotion recognition as well [43, 44].

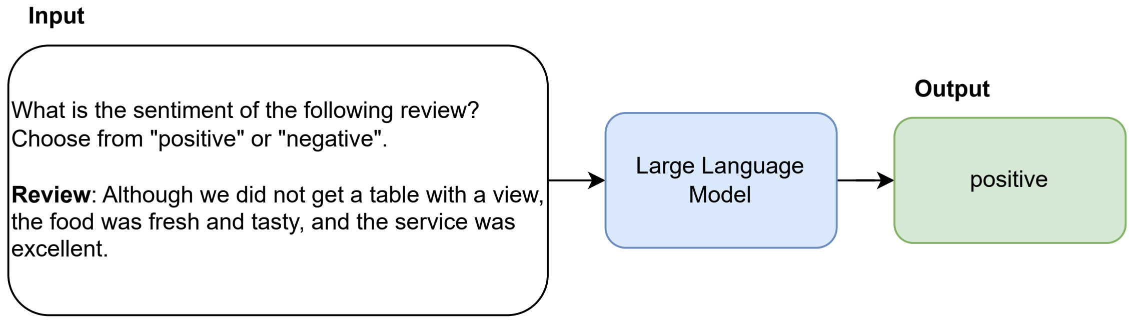
LLMs (Large Language Models) are also widely used for solving various speech-processing tasks. Before LLMs, models like BERT first underwent robust training on large data corpora, but then they were also fine-tuned for a specific task they were used to solve. Large Language Models (LLMs) are unique because, unlike models like BERT, they are initially trained on extensive datasets and then used without further training for specific tasks, which is known as zero-shot usage. However, in some cases, LLMs can be adapted with just a few examples, turning it into a few-shot setting. Figure 5, inspired by paper [45], provides an example of a few-shot learning, because we provide at least one example of how this task has to be solved. In Figure 6, on the other hand, the model did not receive any examples and has to make a guess based only on the task and previously accumulated knowledge.



**Review**: Food was cold, and my waiter was very rude to me.
**Sentiment**: negative.

**Review**: This was the best pizza I have ever tasted! Friendly service, definitely recommend.
**Sentiment**: ?

Figure 5. Example of a few-shot learning.



What is the sentiment of the following review? Choose from "positive" or "negative".

**Review**: Although we did not get a table with a view, the food was fresh and tasty, and the service was excellent.

Figure 6. Example of a zero-shot learning.

As described in [45], beyond what was originally expected of them, LLMs showcased incredible phenomena of emergent abilities. An emergent ability of Large Language Models (LLMs) refers to a new or growing skill or feature that expands their capabilities in tasks like reasoning, engaging in conversations, understanding language nuances, program execution, or instruction following. Emergent abilities are observed in large models, but not in smaller models, and they depend on parameters such as the amount of data, its quality, the number of parameters in the model, and more. These abilities were surprising and not predicted during their initial training, highlighting the model's capacity to learn and adapt beyond its original scope. Work [46] presents the incredible transfer learning abilities of LLMs when adapted for emotion recognition tasks. Similarly, in [47] authors propose a highly competitive with baseline methods novel PromptAV technique based on the LLMs for solving an Authorship Verification task, model proved to be efficient in zero-shot and few-shot settings. Finally, for the task of speaker diarization, as shown in [48], information captured by LLMs enhanced the results of the acoustic-based speaker diarization system.

We can observe that there are countless approaches and their combinations when it comes to using DNNs for speaker recognition, and a vast body of research focusing on this topic has already been conducted. Despite such attention, after completing literature research we deducted that employing content for speaker identification has so far not been researched thoroughly enough. Research also did not reveal any studies where LLMs have been used for speaker identification. Thus, this work has a reduced scope and investigates how using content of the speech along with Pretrained Language Models can improve the results of traditional audio-based speaker identification.

# 3. Supervised speaker identification using text classification

## 3.1 Method

Supervised models are trained specifically for the task that they are expected to solve. The model requires prior training on speech samples from the respective speakers to effectively recognize them. Consequently, before training, we need to prepare training (enrollment) data. In audio-based speaker recognition systems, the actual content in the enrollment (training) data is typically unimportant for the system to recognize the speaker later on. Instead, standardized prompts or phrases are used by every test speaker during the enrollment phase to capture the unique characteristics of the speaker's voice. In this setup, of course, the content of the speech would not carry any relevant identification information. On the other hand in our scenario, where we are testing if the speech content can give extra cues about the speaker's identity, training data must capture natural everyday conversations, like interviews, broadcast shows, etc.

In this work, we experiment with a pretrained BERT-like masked language model RoBERTa. The model is specific to the task of text-based speaker classification. Standard trained model is further fine-tuned to achieve better performance. To do so we prepare speech transcripts of the selected speakers derived from audio files. Each transcript is annotated with a speaker label either by the automatic ASR pipeline or manually.

There are two scenarios when working with speaker identification: closed and open set classification. Closed set classification usually presents fewer challenges, as for each speaker in the test set there is data available in the training set (test and training sets have identical sets of speaker labels). Unlike the closed set, the open set approach better reflects real-life scenarios and presumes that the test set contains speakers that were not seen in the training set (speaker labels set of the test dataset is bigger than speaker labels set of the training dataset). This work offers experimentation of both types: for closed set classification we are employing majorly English VoxCeleb dataset, while for open set identification, we use datasets comprised of Estonian broadcast news, interviews, and recordings from a public opinion festival.

## 3.2 Experiments: VoxCeleb

### 3.2.1 Data

VoxCeleb1, described in [49], is an audio files-based dataset collected via an automated pipeline from YouTube videos of predominantly celebrity interviews. The dataset is compristed of recordings from 1251 unique celebrities, with 55% male and 45% female speakers, predominantly from the USA or the UK. Other less-represented countries include India, France, New Zealand, Canada and Germany. Celebrities in VoxCeleb are a subset of the VGG Face dataset, which comprises individuals at the intersection of the most frequently searched names in the Freebase knowledge graph and the Internet Movie Database (IMDB). The original number of POI was 2,622 identities, but part of the videos were discarded during speaker and face verification stages to minimize labeling errors in the dataset. The dataset contains videos captured in various challenging, multi-speaker acoustic settings, such as red carpet events, outdoor stadiums, quiet studio interviews, speeches to large audiences, and recordings from professional multimedia sources as well as handheld devices. These videos are affected by real-world noise like background chatter, laughter, overlapping speech, and room reverberations. The quality of the recordings varies due to different recording equipment and channel noise. Figure 7 from [49] describes the main stages of the dataset collection pipeline. Stages include compiling the candidate list, downloading the videos, face tracking, speaker verification, and face verification using CNN.

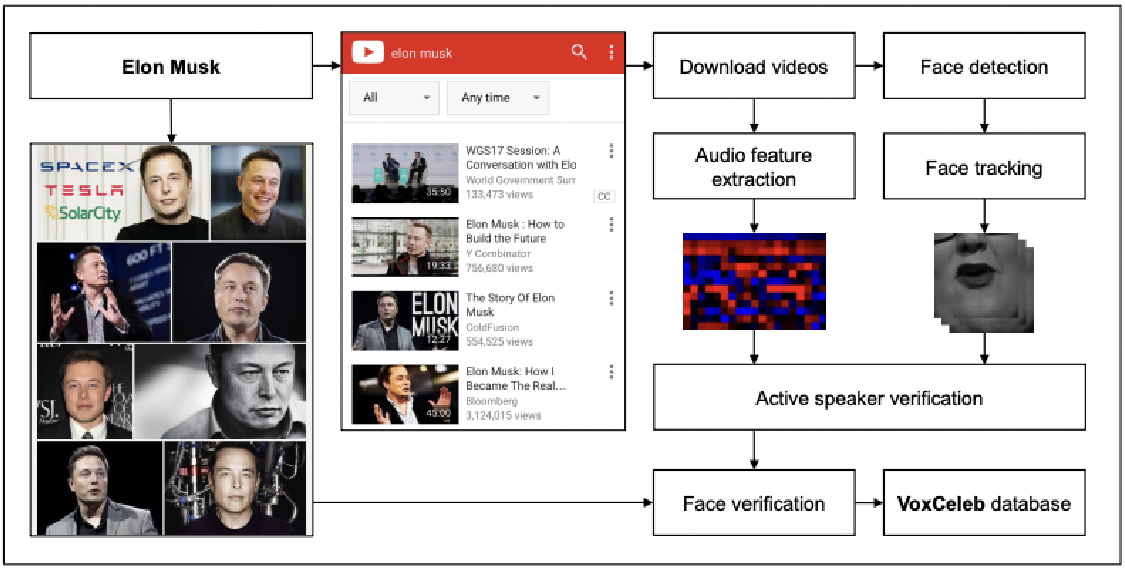


Figure 7. VoxCeleb automated data processing pipeline.

VoxCeleb1 contains official development, test, and train splits for speaker identification. They are constructed so that a single video recording for each speaker is represented in each of the sets. All audio segments extracted from those videos are distributed among the development, test, and train sets, with the biggest part of content dedicated of course to the train set, as for successful model training we require sufficient data. In contrast, development and test sets used for testing model performance can contain relatively short utterances.

The VoxCeleb1 dataset originally contained only raw audio files, and it does not include segment transcripts with the speech content. As in our experiments, we were interested in the text-based speaker recognition experiments, so we transcribed all the datasets using the highly robust speaker recognition system Whisper (whisper-medium), created by OpenAI [50]. During the transcription process, it was revealed that a small subset of VoxCeleb utterances includes non-English speech. As we are using models pretrained on English data in the data processing stage we applied Whisper in translation mode, which successfully translated all speech outside of the English language to English.

In the original dataset, each speaker had multiple segments that belonged to one or more recordings, that were split into 3 datasets. We decided to follow a scenario where the speaker is classified based on all the segments derived from a singles recording. This means that all segments from each recording were concatenated together, resulting in a monolithic recording transcript. In the end, development and test sets contained 1251 test items (recordings), while train sets usually contain more than one recording per speaker. Each speaker in each dataset was assigned a unique identification (ID) for reference. The final data structure mapped the unique speaker ID to a set of one or more recording transcripts, with each recording identified by a recording ID.

During the creation of the audio-based models, we used the SpeechBrain [51] ECAPA-TDNN model trained on Voxceleb1 and Voxceleb2 training data [1] to extract speaker embeddings for audio files. Because the embeddings were created for each segment, when estimating the results we simply averaged the posterior probability distributions for each segment of a single recording. As the last step before model training, each segment was tokenized using a Roberta base tokenizer. Tokenization is essential for converting textual data into numerical form suitable for training neural network models. It breaks down text into tokens that represent words or subwords from a predefined vocabulary, enabling fixed-length input sequences required by neural networks. Tokenization also facilitates padding or truncation of sequences to ensure uniform input lengths, optimizing the efficiency of batch processing during model training.

---

[1] Available at huggingface.co/speechbrain/spkrec-ecapa-voxceleb

### 3.2.2 Experiments

In this part of the work, we are experimenting with the following models: two text-based speaker identification audio models (Naive Bayes and Transformer), and one audio-based model - logistic regression.

In the case of test-based models in both cases, we found it beneficial to apply data augmentation to the training data: we performed a sentence dropout, meaning that if the training sample contained several sentences, a random sentence was removed and this new sample was added to the dataset along with the original one. We repeated the process as many times as many sentences as there were in the sample.

Although Naive Bayes is a relatively simplistic model based on the assumption that the classes are independent of each other, it can surprisingly well compete with other more complex models. We trained the Naive Bayes model using unigram word features. Initially, the model achieved an accuracy of 2%, but applying data augmentation increased accuracy to 9% and stop word removal to 12%. Stop words are frequently occurring words like 'a' or 'the', but they do not carry meaningful information. Removing stop words can let the model focus on more significant classification information.

When experimenting with a Transformer model we utilized a pretrained case-sensitive version of the RoBERTa (large) [25] model as the starting point. It was finetuned for text classification in a standard way, using the augmented training data. We selected a learning rate of 1e-5, weight decay of 0.01, and 10 training epochs as base parameters for training. Augmentation of data helped to improve the model accuracy from 14.1% to 21.3%.

We also wanted to see how the RoBERTa model would perform if we reduced the number of possible speakers to ten most likely candidates. Employing the audio-based model, we identified each segment's top 10 most probable guesses by selecting the ten highest probabilities from the posterior probabilities array. Further, when we receive the posterior probabilities from the RoBERTa model we selected only those corresponding predictions recommended by the audio model and then further chose the highest-probability prediction as normally. This process yielded a 48.1% accuracy rate.

Subsequently, we also wanted to receive a model trained on the audio ECAPA-TDNN embeddings. We selected a simple multinomial logistic regression model that was trained on the training split of Voxceleb1. Logistic regression is a statistical method used for predicting the probability of a binary outcome based on one or more predictor variables.

Table 2. Results on VoxCeleb1 dataset with supervised models.

| Model | Accuracy (%) |
|---|---|
| Audio-based: EPACA-TDNN | **99.8** |
| Text-based: Naive Bayes | 12.2 |
| Text-based: RoBERTa | 21.3 |
| RoBERTa (on top 10 audio predictions) | 48.1 |

### 3.2.3 Results and analysis

The results of the experiments are presented in Table 2. As expected, the audio-based model achieved a staggering 99.8% accuracy. This corresponds to only two identification errors out of 1251 recordings. RoBERTa combined with the audio model has achieved 48.1% accuracy, while the standalone RoBERTa model has only reached 21.3%. Naive Bayes was proved to yield the lowest accuracy of 12.2%. Because the audio-based model is extremely accurate, we didn't try fusing the predictions of audio and text-based models, as the results would not be statistically significant. The results show the importance of data processing, as simple train data augmentation has improved the results of the RoBERTa model from 14.1% to 21.3%.

## 3.3 Experiments: Estonian broadcast and public debate speech

### 3.3.1 Data

**Data collection**

Table 3. Training, development and test data for Estonian experiments. For each source, the number of recordings and the total amount in hours is given.

| | Training | Dev | Test |
|---|---|---|---|
| News clips | 10585 / 458 h | | |
| Evening radio news | 7109 / 1978 h | 889 / 316 h | 889 / 326 h |
| Talk shows | 1236 / 1000 h | 154 / 134 h | 154 / 136 h |
| Opinion Fest. | | | 51 / 75 h |

In experiments with the Estonian language, we employed three datasets that were collected from radio news, radio talk shows, and recordings of a public opinion festival (*Arvamusfestival*). Training data and the segment of test data consist of episodes scraped from the Estonian Public Broadcasting (*ERR*) archive. The majority of the archived episodes feature manually annotated names of the speakers that appeared in them. For training and evaluation of the model, we utilized the recordings of brief radio news snippets, a popular evening radio news program called (*Päevakaja*) (see Appendix A), and an informal radio debate program (*Reporteritund*) that focuses on politically important issues of the day. Programs showcase a large number of speakers from various backgrounds, and conversations cover a wide range of topics, making them a suitable source of data for the experiments. More details of the Estonian data are presented in Table 3. Most episodes were recorded between the years 2004 and 2022, but a few segments are older dating back as far as 1959. The evening news and debate program subsets were further divided between all three training, development, and test sets, while Opinion Festival data only appeared in the test dataset. To evaluate more relevant data, we first sorted each subset by date in ascending order. We extracted the development and test items from the end of the created lists along the time axis, effectively extracting more recent recordings.

As this task is an open-set classification task, we included recorded sessions of the 2021 Estonian Opinion Festival (*Arvamusfestival*) in the test dataset to test out-of-domain performance. It is a debate organized for people from different communities who come together to engage in discussions about civil matters. The setting of the festival session can be seen in 8, found on the Instagram page of the festival. The Festival is structured around 90-minute discussion sections split by social groups, that are conducted by 5 experts who discuss the topic related to their area of expertise. For moderating the discussion and managing the work with the audience each panel has a designated person. The discussions are transcribed by hand and annotated with the full names of the speakers that could be inferred by the annotator from the introductions.

### Data clustering

In order to illustrate the set of speakers in the Estonian dataset further, we prepared the depiction of the clusters in which speakers were classified. We first selected the last hidden state of the model that was previously trained on the train dataset, and then we took the mean across the sequence dimension, thus creating embeddings representing speaker characteristics encoded within the speech data. Later we employed dimensionality reduction techniques. Principal Component Analysis (PCA) is initially utilized to reduce the high-dimensional embeddings into a lower-dimensional space. PCA, a linear transformation method, effectively captures the most significant components of the data, aiding in compression and simplification without significant loss of information. Following PCA, clustering is performed using K-means, an unsupervised learning algorithm, to partition the speaker embeddings into distinct clusters. We set the number of clusters parameter t0 10. Subsequently, t-Distributed Stochastic Neighbor Embedding (t-SNE) is applied for further dimensionality reduction, specifically tailored for visualization purposes. t-SNE emphasizes the preservation of local relationships, allowing for the embedding of high-

Figure 8. ArvamusFestival session.

dimensional data into a two-dimensional space, thereby enabling intuitive visualization of speaker clusters.

The scatter plot illustrates speaker embeddings, revealing how speakers are grouped based on the topics of their speech. Each point on the plot represents a speaker, with colors denoting clusters identified by the K-means algorithm. This visualization aids in discerning patterns among speakers within the dataset, providing insights into their domains of interest. The clusters depicted in Figure 9 represent different professional domains, although it's worth noting that they may not necessarily correspond directly to the profession of the speaker. Rather, speakers are grouped based on the similarity of topics discussed, with speakers addressing related topics positioned closer to each other on the cluster map. We identified 13 clusters of speakers, encompassing domains such as healthcare, law enforcement, and journalism, etc. For instance, Kadri Simson, currently serving as a European Commissioner for Energy, was correctly assigned to the national government cluster 10, alongside other representatives of similar categories. Notably, Kadri Simson is positioned closely to politician Jaanus Karilaid, as they were both affiliated with the same political party (Keskerakond, Central Party). Similarly, Martin Herem, an Estonian general and current Commander of the Estonian Defence Forces, is clustered near Leo Kunnas, a former Estonian military officer, and Neeme Väli, an Estonian Major General of the Estonian Defence League, within cluster 11 of military personnel. This scatter plot proves that speakers can be discriminative from one another based solely on their topics, but it also shows that many samples overlap, meaning that the topics discussed are very similar, which could confuse the model based solely on content.

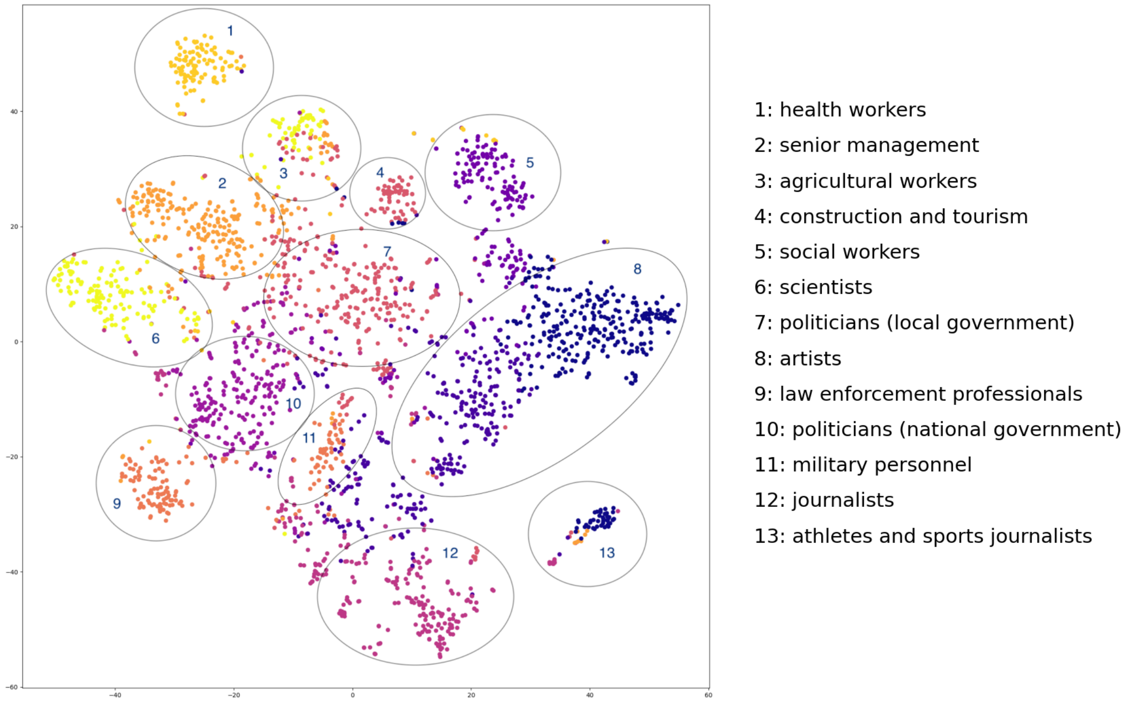


Figure 9. Visualization of Estonian Speaker Embeddings: Clustering of Speakers Based on Speech Characteristics.

## Data preprocessing

To create speaker embeddings for Estonian data we turned to earlier works. Originally each recording from the radio segments (which include news, talk shows, and debates) was annotated with information about the speakers present throughout the entire duration of that recording, rather than being segmented and labeled at specific time intervals within the recordings. Recent work [52] improves the weakly supervised training method that was originally proposed in [53] to train the audio-based speaker recognition models. The previous model used i-vectors for speaker classification. At the same time, in the new approach, ECAPA-TDNN [54] model was used, with the output layer specific to Estonian data that was randomly initialized. The backbone was then finetuned using a smaller learning rate. The model includes speakers who appear at least 10 times in the dataset, totaling 2591 names. The model with 2591 persons has a coverage of 73.0% the the radio news test set, 63.1% on the talk show test set, and 39.9% on the opinion festival test set.

Further, we proceeded to retrieve text data for content-based experiments, Estonian data underwent automatic diarization and transcription using the Estonian transcription system outlined in [55]. The speech recognition models utilize XLS-R-1B wav2vec2.0 models [56], fine-tuned for Estonian ASR with 761 hours of manually transcribed speech, primarily sourced from conversational broadcasts. The system achieves a word error rate (WER) of approximately 8% on conversational broadcasts, with even lower rates on broadcast news.

Taking into account that the training data has weak labels, a weakly supervised audio-based speaker identification model was employed to label the (unnamed) speakers in the training data. This process is called self-labeling. Then all transcribed speech segments corresponding to the labeled speakers were used to train the text-based model. This implies that the text-based model's ability to recognize speaker names aligns with that of the acoustic model, as the acoustic model categorizes all unrecognized names under a unified "unknown" speaker class.

Evaluation is performed on the recording level because the speakers in the radio test data are annotated at the recording level. This means that there is no predefined mapping between diarized speakers and the speaker names mentioned in the radio shows. Models still classify individual speakers identified by the speaker diarization process, but additionally, all speaker names detected within each show are collected, forming a set, that is then compared to a reference set of speaker names associated with that particular show. The comparison is made using precision and recall metrics. The thresholds for precision are adjusted to ensure that the identified speaker names are correct with a high degree of confidence, aiming for at least 95% precision. This is important because incorrect speaker identification could lead to undesired outcomes from an application perspective.

### 3.3.2 Experiments

The Estonian text classification model was trained using a method similar to that of English data. Specifically, we fine-tuned the case-sensitive version of the multilingual XLM-RoBERTa model for a single-label sequence classification task, as described in [57]. Training parameters, including a learning rate of 1e-5, weight decay of 0.01, and 10 epochs, remained consistent. Although we employed a data augmentation technique similar to that used for VoxCeleb, it did not result in improved model accuracy.

To integrate the text-based model with the audio-based model, we obtained text embeddings for all speech turns in the textual training data. This involved extracting the output of the last hidden layer of the XLM-RoBERTa model for each text, selecting the vector corresponding to the first $[CLS]$ pseudo-token. Using these embeddings, we trained a speaker recognition model by reducing the dimensionality of the embeddings to 150 after normalization and implementing a generative classifier based on the PLDA paradigm.

To enhance speaker identification accuracy, especially when the audio-based model shows uncertainty, we devised a strategy leveraging both audio and text-based models. Typically, the audio-based model confidently predicts speakers with a probability above 0.95 or assigns them to the "unknown" class, but its certainty diminishes in intermediate cases.

We focused on refining identification in the scenarios where certainty falls in the intermediate range, which means that the model shows uncertainty. In such cases, the text-based model offers supplementary insights to either confirm or challenge the audio-based model's initial identification. We processed the development and test datasets with the audio-based model, extracting predictions exceeding 1% probability (excluding "unknown" speakers), keeping in mind that for many diarized speakers, there could be several such predictions. Then, we calculated the log-likelihood ratio scores from the text-based LDA/PLDA model for each speaker and speech turn pair.

Combining the audio-based model's probabilities and text-based model scores, along with binary reference labels indicating speaker presence, we trained a logistic regression model. This model was used to generate final speaker identification predictions for the test data. Figure 10 describes the interpolation of the models.

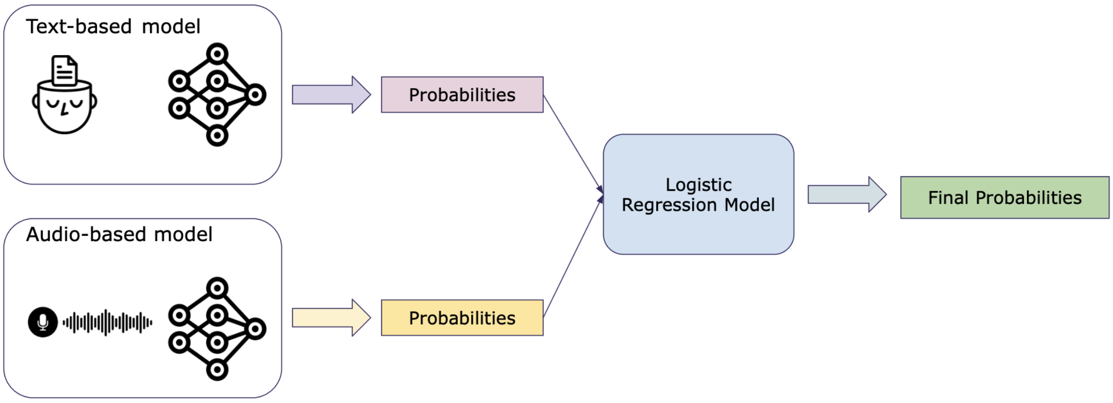


Figure 10. Interpolation of the RoBERTa and audio models.

### 3.3.3 Results and analysis

The speaker identification results for three test sets are presented in Table 4. It's evident that the audio-based model achieves high precision on its own, while the text-based model lags significantly behind. This outcome was expected, as the text-based model primarily identifies speakers with high confidence only when they explicitly introduce themselves, a common practice in radio news broadcasts. Despite its limitations, integrating the text-based model with the audio-based model results in a modest improvement in precision on the news and opinion festival datasets compared to using the audio-based model alone.

Table 4. Speaker identification precision (P) and recall (R) rates of different models on Estonian test sets.

| | News | | Talkshows | | Op. festival | |
|---|---|---|---|---|---|---|
| | P(%) | R(%) | P(%) | R(%) | P(%) | R(%) |
| Audio-based | 98.4 | 71.7 | 94.7 | 64.2 | 96.8 | 26.7 |
| Text-based | 81.8 | 20.8 | 85.8 | 16.2 | 11.1 | 0.3 |
| Audio + text PLDA | 98.5 | 71.8 | 94.7 | 64.8 | 98.9 | 26.4 |

This improvement is particularly notable in cases where the audio-based model assigns a moderate posterior probability (e.g., 50%) to a speaker, and the text-based model effectively clarifies these predictions in the correct direction.

For example, the audio model was hesitant when it attempted to recognize a speaker from the evening news "Paevakaja". The true speaker identity is Olari Elts, an Estonian conductor, but the audio model only gave this label 0.333, which was not enough to label the segment correctly. In this case, the text-based model was able to overturn that decision, based on his interview transcript: *"Algusest peale soov korraldada Hiiumaal kõrgetasemelist klassikalist muusika, festivali, mis ei ole nii-öelda ainult Hiiumaa festival, vaid selline nagu maailma mastaabis, et siia tulla tahaks kuulama ka nii-öelda välisriikidest. Teiseks, et Erkki, Sven Tüüri looming ja eriti tema orkestrilooming kindlasti kõlaks tema kodusaarel. Ja kolmandaks on siis see, et toetada Hiiumaa vanimat kirikut, pühalepa kirikut ja seal sees olevat orelit."*

However, in the talk show domain, where speakers typically talk for longer durations, the audio-based model tends to make more confident decisions, and the text-based model does not contribute to improved results.

# 4. Zero-shot speaker identification with LLMs

## 4.1 Method

This section investigates zero-shot speaker identification using large language models. We utilize the OpenAI API to transmit transcripts of speakers' utterances, as in the case of VoxCeleb, or transcripts of recordings, as in the Estonian dataset, and request the model to identify these speakers. This process constitutes unsupervised identification, as the model has not been previously trained on the train dataset. However, given that the GPT model was originally trained on extensive data sourced from the internet, it is probable that it has encountered some relevant information previously, such as interviews and online articles. Since LLMs are not specifically fine-tuned for this task, as no training samples were provided to the model, it is more accurate to characterize this scenario as zero-shot inference.

In this work we use OpenAI's GPT-3.5 and GPT-4 [22] models for experiments. Despite the recent release of several freely available large language models (LLMs), none of them have substantial capabilities for processing Estonian data.

It has to be noted that GPT identification results will differ drastically for the English and Estonian datasets due to the nature of the data in them. In the case of VoxCeleb1 dataset, it includes exclusively speech of the celebrity, meaning that actual names of the celebrities appear in the dataset extremely rarely. On the other hand, the Estonian dataset includes an introduction from the hosts, meaning that there are often cases when the name of the target speaker is explicitly mentioned, which allows the model to identify speakers more accurately.

## 4.2 Experiments: VoxCeleb

### 4.2.1 Data

In the experiments conducted in this chapter, we utilize the pre-existing VoxCeleb1 dataset, transcribed from supervised text-based experiments. The VoxCeleb test data is prepared by aggregating individual utterances from each test video into a cohesive string.

Figure 11 illustrates the distribution of utterance lengths in the VoxCeleb1 training dataset

Table 5. Example predictions of the GPT-4 model with true speakers.

| Input | Top 5 hypothesized speakers | True Speaker |
|---|---|---|
| It's just like, you know, everyone nowadays goes around screwing everyone. He just likes her and doesn't want to be a... I guess he doesn't really want to cheapen his relationship with her. I mean, he doesn't want to mess it up. He thinks she's the love of his life, so he wants to do it right. I don't know. I mean, I guess when Kristen brings a lot of herself to Bella, and I'm not sure if, like, when I read the books and got to know Kristen, I was like, she's not really like the Bella I imagined in the books. She seems much more of a damsel. But Kristen's quite a lot tougher, and it comes across in the movies like that, and a lot more. | **Robert Pattinson**<br>Taylor Lautner<br>Peter Facinelli<br>Kellan Lutz<br>Jackson Rathbone | Robert Pattinson |
| I would start making pop music and I would stop writing smart lyrics or I would stop writing. No, not at all. And that's why when I go online and I go on Instagram and I see, you know, a post from Emma who lives in Philadelphia and she's talking about how her day was at school that day. She's oversharing, and she's over-emotional, or she might be crazy, or watch out, she's... and all the themes that used to be main factors in my music. | **Taylor Swift**<br>Ariana Grande<br>Selena Gomez<br>Demi Lovato<br>Miley Cyrus | Taylor Swift |
| Very excited and pretty shocked because I didn't have to audition, which I thought was very weird. So yeah, and I think the drama of it all mixed in with the humor and the romance kind of makes for a really exciting movie. I pretty much didn't really emotionally prepare that much for the movie except for the fact of wondering what it would be like if my mom was taken. Otherwise, it was just how would Lily react as Clary in these situations. She is pretty much every normal girl out there that finds out that they're not normal. So, what would I do? So basically, there are so many times I look on the screen and I just... it's not about vampires versus werewolves... | **Lily Collins**<br>Emma Watson<br>Jennifer Lawrence<br>Kristen Stewart<br>Shailene Woodley | Lily Collins |
| I mean, there's something really cool about a mythology being able to change, having creative license to change a mythology to adapt to modern times. I think what's great about the old classic Draculas was that synthesis of that. I always found that really amazing. Plus, Willem Dafoe is one of the best actors. I was a bad boy growing up, but I was the bad boy who was in disguise because I was the favorite. Or in the police force, do you do anything bad, but I was kind of a... | Johnny Depp<br>Robert Pattinson<br>Leonardo DiCaprio<br>Brad Pitt<br>Tom Cruise | Ian Somerhalder |

per speaker. The lengths vary, with the majority falling between 4,000 and 20,000 characters. The shortest utterance contains 2,475 characters, while the longest consists of 147,337 characters. Unlike with Estonian data, the clustering of speakers in the VoxCeleb1 dataset yielded insignificant outcomes. This is primarily attributed to the predominant occupation of the speakers, who are overwhelmingly actors. Consequently, deriving clusters based on the topics of speech proves unfeasible.

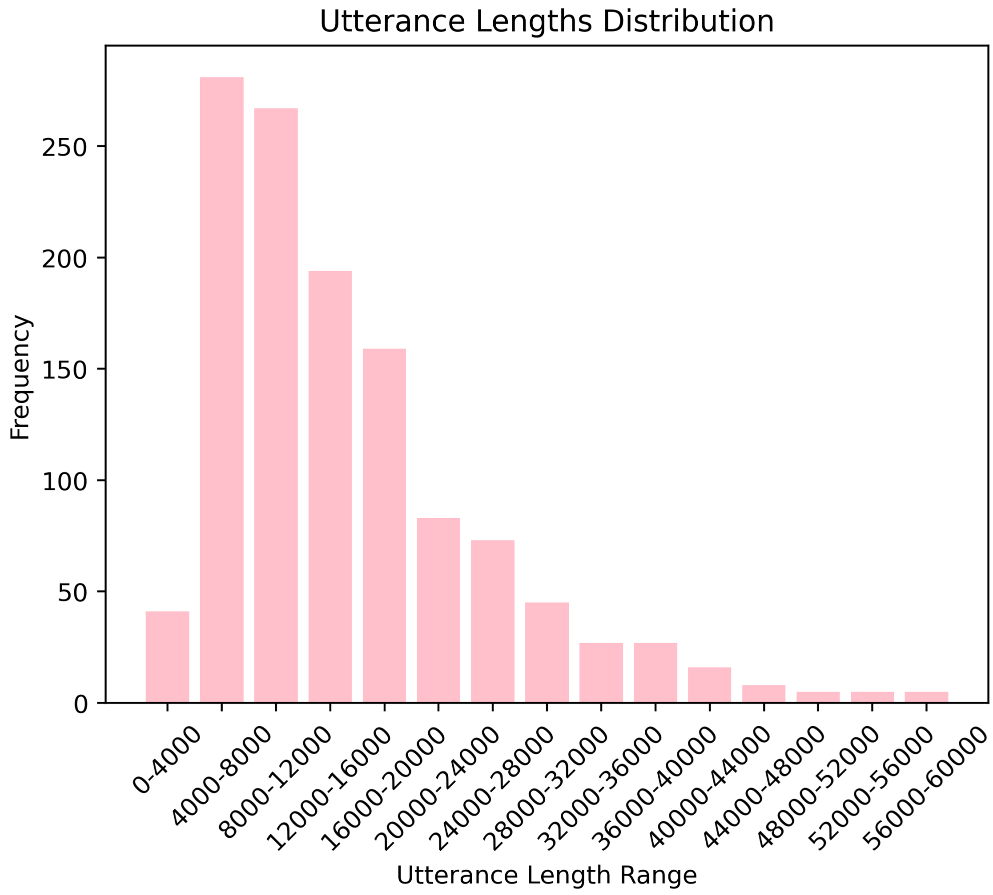


Figure 11. Length of utterances of the transcribed VoxCeleb1 dataset.

The GPT model receives this concatenated transcription of the recording text for a speaker, such as a VoxCeleb interview, without any additional context or clues regarding the speaker's identity. Unlike supervised closed-set approaches, in the majority of experiments, the model is not trained on the train dataset, nor does it possess a predefined pool of identities to restrict the search. Instead, the model relies solely on its original training data. In one experiment, we provided the model with a set of 10 possible candidates from which to select. The candidate pool was selected based on the predictions of the audio-based model.

### 4.2.2 Experiments

In this section, we examine the efficacy of LLMs in speaker identification under zero-shot conditions. We employed the OpenAI API interface to iteratively query a GPT LLM, utilizing prompts structured in the following format:

> Here are some interview segment transcripts from a certain celebrity. Who do you think it might be? Please provide the top 10 guesses. Present the result as a JSON-formatted list of lists, for example: [[First Name, Second Name], [First Name1, Second Name1]]. In case you cannot identify an individual, please provide your best 10 guesses; otherwise, present an empty list.
>
> [Transcript of the utterances]

As a response, we asked the model to generate a JSON array containing the list of the list of 10 possible guesses from most to least likely that we later cross-referenced with the true celebrity name in order to evaluate the model performance. The example of the complete flow is presented in 15. You can see that the model can successfully identify individuals with a significant online presence, especially if additional hints are provided, such as the name of the related company or a movie, but in some cases, the context is not enough. In this example, the model could not identify the speaker Clive Owen successfully (S4).

We performed the experiment on both OpenAI's GPT-3.5 (gpt-3.5-turbo-0613) and GPT-4 (gpt-4-0613).

In the second part of the experimentation, we leveraged the highly accurate audio-based speaker identification model that we trained in the previous chapters. We selected the top 10 most likely celebrities per each segment that we sent the GPT model (celebrities that corresponded to the highest posterior probabilities). We provided the model with this list of top 10 celebrities in random order and instructed it to select the correct name from this list. As opposed to the earlier experiments, this experiment is a closed-set identification task.

### 4.2.3 Results and analysis

Table 6 depicts the results, where we evaluated both Top1 accuracy and Top10 accuracy. The Top1 accuracy reflects how well the model managed to effectively identify the celebrity as its first guess (whether the correct celebrity's name placed first in the list). However,

Table 6. Results on VoxCeleb data with GPT models.

| Model | Accuracy (%) |
|---|---:|
| GPT-4 Top1 | 22.5 |
| GPT-4 Top10 | **31.3** |
| GPT-3.5 Top1 | 3.1 |
| GPT-3.5 Top10 | 5.9 |

Top10 accuracy signifies how successfully the model could include the name of the correct celebrity within the entire list of predictions (whether the correct celebrity's name was found anywhere in the list).

Results of the experiments where the top 10 probable candidates are provided to the model are shown in Table 7. It can be seen that in 80.8% of the cases, GPT-4 picked the correct celebrity from the provided set, while GPT-3 only guessed right in 58.5% of the cases.

Table 7. Results on VoxCeleb data with GPT models, when provided the 10 most probable speakers from the audio-based model.

| Model | Accuracy (%) |
|---|---:|
| GPT-4 | **80.8** |
| GPT-3.5 | 58.5 |

Table 5 presents examples of predictions made by the GPT-4 model using the provided input. The analysis reveals instances where the model accurately identifies a speaker as the primary prediction, while in other cases, such as Ian Somerhalder, the identification process is less straightforward. Notably, we observed that utterances containing references to television shows or associated characters notably aid in speaker identification.

## 4.3 Experiments: Estonian broadcast and public debate speech

### 4.3.1 Data

The distribution of utterance lengths in the Estonian training dataset per speaker is depicted in Figure 1. Similar to the VoxCeleb1 dataset, the lengths of utterances exhibit variation, with most utterances falling below 20,000 characters. The shortest utterance comprises 43 characters, while the longest recorded speaker utterance consists of 3,352,047 characters. The portion of the dataset attributed to "Unknown" speakers accounts for 55,470,801 characters.

When conducting experiments with the Estonian dataset, we furnish the model with

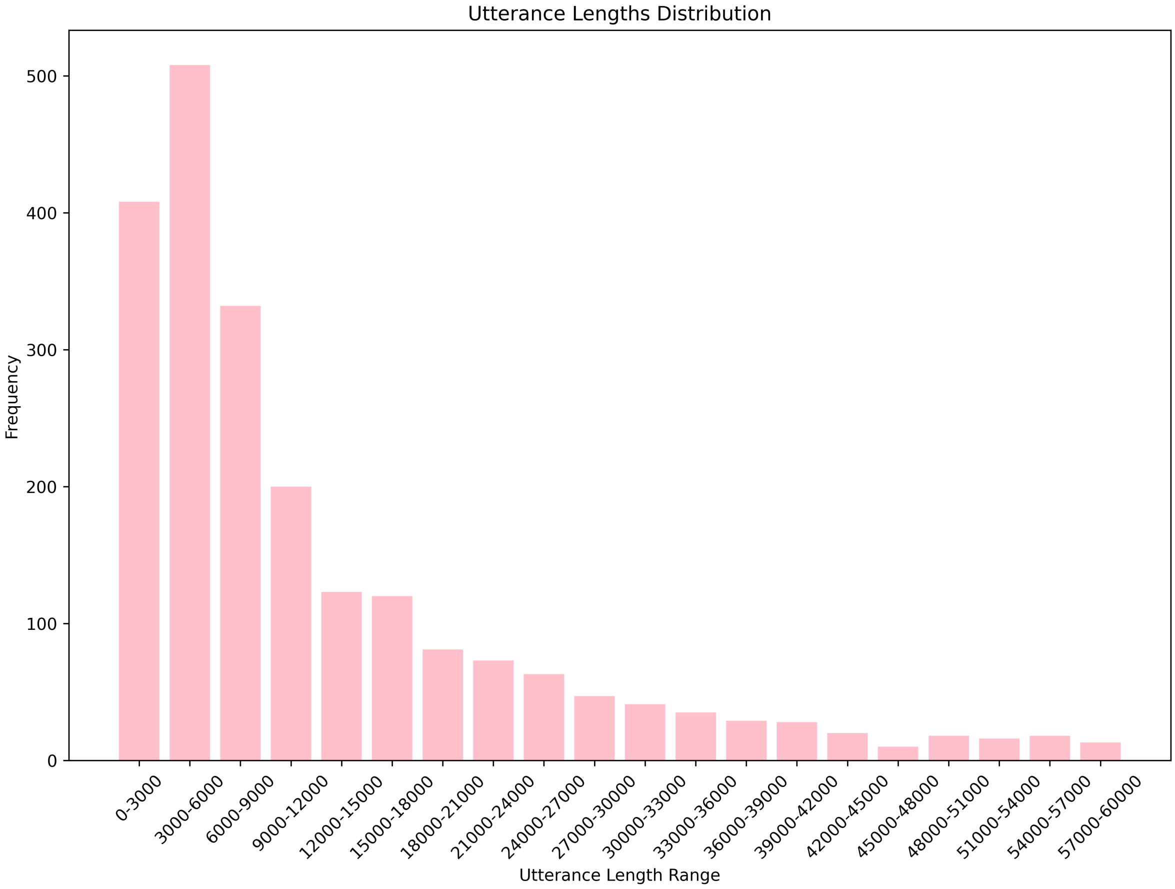


Figure 12. Length of utterances of the transcribed Estonian dataset.

the complete transcript of the recording. This transcript amalgamates all speaker turn transcriptions along with their respective speaker labels, facilitated by robust transcription and diarization systems. Consequently, the model gains access to a broader spectrum of information, including the various ways the speaker is addressed, as highlighted in previous literature [27, 28, 29, 30, 31, 32, 33], in addition to the speech content.

### 4.3.2 Experiments

Since we work with Estonian data, our query for the model has to be in Estonian as well. Translated to English prompt is the following:

> You are an expert in Estonian public figures. You will be given an automatic transcription of the news or talk show, complete with speaker codes. Try to guess which persons are speaking in the program and also find the connection between the speaker codes and names. Output the result using JSON. JSON format example: "code: "name". If you don't know the name, write "Unknown" instead of the name. Don't take too many risks, accuracy is

> more important to us than yield. If you are not particularly sure of the match, write instead "Unknown". Names may be incorrectly transcribed, use your background knowledge to correct them if necessary.

The workflow depicted in Figure 16 illustrates the entire process: model input (instructions to the model), as well as input transcriptions. Due to the extended duration of opinion debates, which typically span 90 minutes, the transcripts tend to exceed 32k tokens. At the outset of our experiments, the GPT-4 model had a context length limit of 8k tokens. Consequently, we divided each transcript into multiple segments for processing. In each segment, the prompt also incorporates the mappings of speaker codes to speaker names, which are predicted from the preceding segment, along with the relevant instructions.

To cut down on costs with the OpenAI API, we tested our system using random sets of 20 shows from both broadcast news and talk shows. This is why the results from the audio-based model might differ compared to what you see in Table 8. But lately, the costs of using the OpenAI API have gone down a lot. Now, processing a 90-minute conversation transcript only costs around $0.30.

Figure 13 described the flow of the interpolation of the GPT4 and audio models. The audio model has the same architecture as described in the Chapter 3. Due to the high accuracy of the audio model, we generally use its predictions, and only in cases where the model could not confidently generate a prediction we use the prediction generated by the GPT4 model.

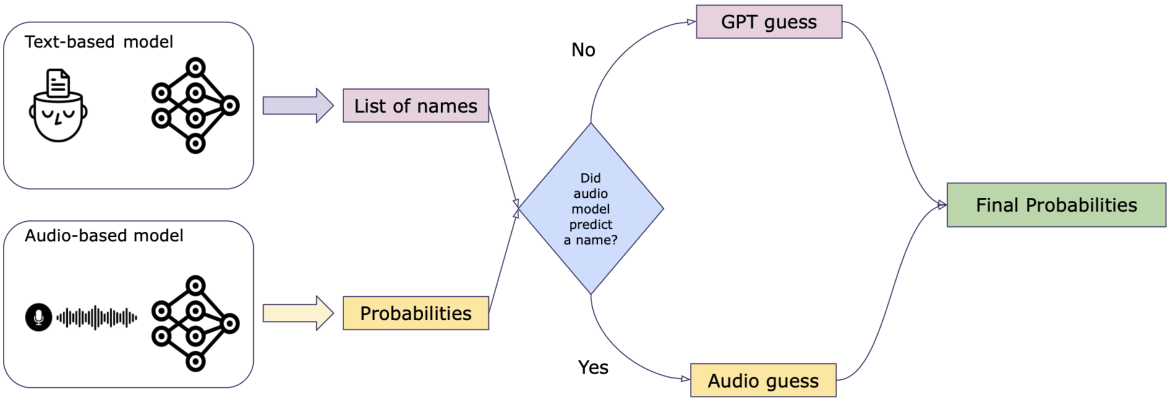


Figure 13. Interpolation of GPT4 and audio models.

### 4.3.3 Results and analysis

The results of speaker identification using LLMs are presented in Table 8. Precision and recall rates for LLM-based systems are computed considering a 1-character Levenshtein

Table 8. Precision and recall rates of LLM-based speaker identification on the Estonian test sets. We compared OpenAI's GPT3.5 (with 16k token context size) and GPT4 (with 128k context size).

| | News | | Talkshows | | Op. festival | |
|---|---|---|---|---|---|---|
| | P (%) | R(%) | P(%) | R(%) | P(%) | R(%) |
| Audio-based model | 99.6 | 69.9 | 95.9 | 52.2 | 96.8 | 26.7 |
| GPT 3.5 (16k) | 97.1 | 10.6 | 100.0 | 47.3 | 90.7 | 28.4 |
| GPT4 (128k) | 97.5 | 71.4 | 100.0 | 97.8 | 97.1 | c |
| Audio + GPT4 | 99.0 | 89.9 | 97.8 | 97.8 | 96.9 | 73.6 |

distance with respect to the reference names. For opinion festival transcripts, instances, where LLMs predict speaker names without surnames (common as speakers, are often introduced without surnames) were excluded from the analysis.

GPT-4 (gpt-4-1106-preview) demonstrates notable precision and recall rates across all test sets. Particularly on talk shows and public debate data, it significantly outperforms the audio-based model. This discrepancy is largely due to the audio-based model having relatively low coverage of names on these datasets, whereas the LLM infers all names directly from the transcripts. Furthermore, it's apparent that GPT-4 performs substantially better on this task compared to GPT-3.5 (gpt-3.5-turbo-16k-0613).

The final row in Table 8 corresponds to a combined system. This system integrates predictions from the audio-based model for each speaker code, if available, and utilizes GPT-4-based name hypotheses for speakers not identified by the audio-based model.

In Appendix 2, we provide an additional example of the full news recording transcript based on which GPT4 model attempts to make a prediction. It can be seem how GPT4 model can successfully derive the names of speakers from the text.

As previously mentioned, we employed a one-character forgiveness distance when comparing speaker names proposed by LLMs with those of the reference speakers. This adjustment is necessary due to occasional small errors in person names found even within the reference transcripts, particularly for first names that may be ambiguously written in the language. The issue becomes more pronounced when the LLM relies on ASR-generated transcripts. Table 9 illustrates precision and recall rates on opinion festival recordings, comparing the use of ASR-generated transcripts versus reference transcripts, with increasing forgiveness distances. As anticipated, performance on ASR-generated transcripts is slightly lower, although not substantially so, particularly considering the relatively noisy nature of the recordings, which poses challenges for the ASR system.

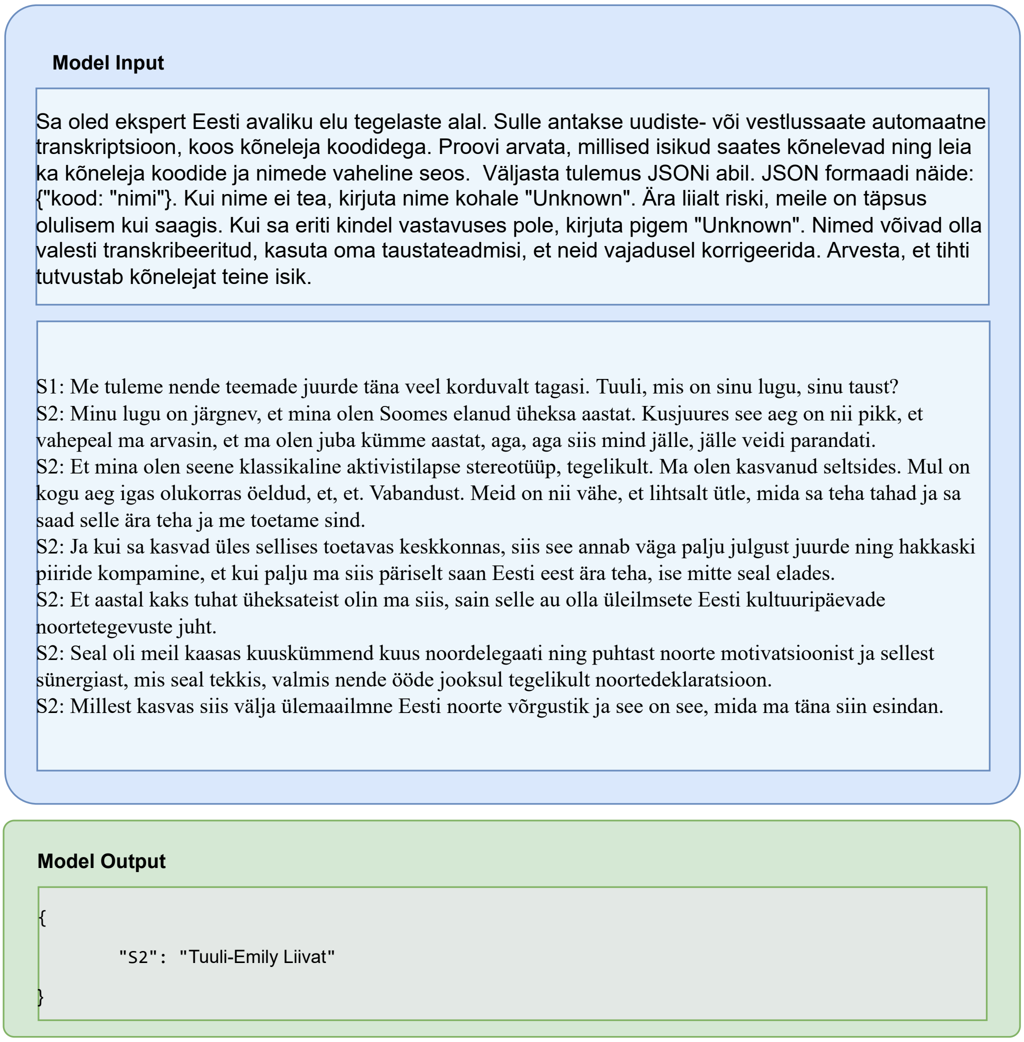
Table 9. Speaker identification precision and recall on Estonian opinion festival transcripts with GPT-4, based on ASR transcripts *vs* reference transcripts, and with increasing name comparison edit distance.

| | ASR | | Reference | |
|---|---|---|---|---|
| Edit distance | P(%) | R(%) | P(%) | R(%) |
| 0 | 91.7 | 64.5 | 95.1 | 68.0 |
| 1 | 94.2 | 66.3 | 97.1 | 69.5 |
| 2 | 95.8 | 67.4 | 97.5 | 69.8 |

During our analysis of GPT-4 generated speaker names for the opinion festival data, we observed instances where the model proposed names for speakers who were unnamed in the reference transcripts. These occurrences often arose when a speaker from the audience introduced themselves using only their first name before proceeding to ask a question or contribute to the debate. Since the human annotator did not possess the full name of the speaker, they remained unnamed in the reference transcripts. However, GPT-4 was able to infer the full name of the speaker based on the context of the question, particularly if the individual was a recognized spokesperson on the given topic or a local journalist covering relevant issues. We conducted thorough analyses of such cases by seeking speech samples for these specific speakers from online sources and manually comparing them to the corresponding speech segments in the test data. In the majority of cases, we found the inferred names to be correct.

An illustrative example is presented in Figure 14. Here, the target speaker is mentioned as "Tuuli," yet the output from the GPT-4 model provides the full name "Tuuli-Emily Liivat." This discrepancy in performance can be attributed to Tuuli-Emily's significant online presence and her involvement in the Estonian community in Finland. It is plausible that the model encountered data related to her during pretraining, thus enabling it to make more informed predictions about her identity.

Another noteworthy capability of the GPT-4 model is demonstrated in Figure 16, where it successfully lemmatizes names. Despite the source text addressing "Indrek Kiisler" as "Indrek Kiislerile," the model accurately specifies the name as "Indrek Kiisler."



**Model Input**

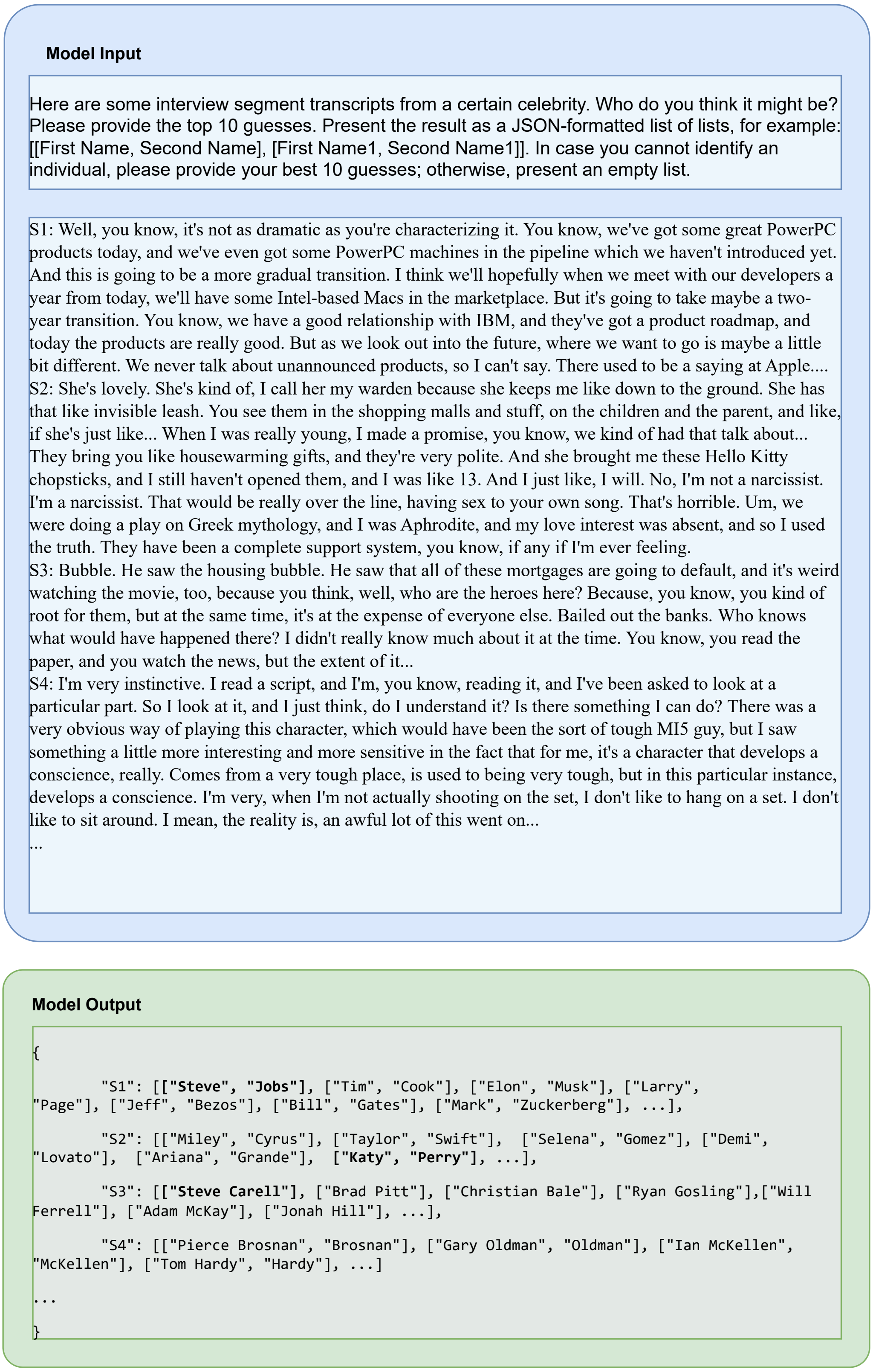
Sa oled ekspert Eesti avaliku elu tegelaste alal. Sulle antakse uudiste- või vestlussaate automaatne transkriptsioon, koos kõneleja koodidega. Proovi arvata, millised isikud saates kõnelevad ning leia ka kõneleja koodide ja nimede vaheline seos. Väljasta tulemus JSONi abil. JSON formaadi näide: {"kood: "nimi"}. Kui nime ei tea, kirjuta nime kohale "Unknown". Ära liialt riski, meile on täpsus olulisem kui saagis. Kui sa eriti kindel vastavuses pole, kirjuta pigem "Unknown". Nimed võivad olla valesti transkribeeritud, kasuta oma taustateadmisi, et neid vajadusel korrigeerida. Arvesta, et tihti tutvustab kõnelejat teine isik.

S1: Me tuleme nende teemade juurde täna veel korduvalt tagasi. Tuuli, mis on sinu lugu, sinu taust?
S2: Minu lugu on järgnev, et mina olen Soomes elanud üheksa aastat. Kusjuures see aeg on nii pikk, et vahepeal ma arvasin, et ma olen juba kümme aastat, aga, aga siis mind jälle, jälle veidi parandati.
S2: Et mina olen seene klassikaline aktivistilapse stereotüüp, tegelikult. Ma olen kasvanud seltsides. Mul on kogu aeg igas olukorras öeldud, et, et. Vabandust. Meid on nii vähe, et lihtsalt ütle, mida sa teha tahad ja sa saad selle ära teha ja me toetame sind.
S2: Ja kui sa kasvad üles sellises toetavas keskkonnas, siis see annab väga palju julgust juurde ning hakkaski piiride kompamine, et kui palju ma siis päriselt saan Eesti eest ära teha, ise mitte seal elades.
S2: Et aastal kaks tuhat üheksateist olin ma siis, sain selle au olla üleilmsete Eesti kultuuripäevade noortetegevuste juht.
S2: Seal oli meil kaasas kuuskümmend kuus noordelegaati ning puhtast noorte motivatsioonist ja sellest sünergiast, mis seal tekkis, valmis nende ööde jooksul tegelikult noortedeklaratsioon.
S2: Millest kasvas siis välja ülemaailmne Eesti noorte võrgustik ja see on see, mida ma täna siin esindan.

**Model Output**

```
{
        "S2": "Tuuli-Emily Liivat"
}
```

Figure 14. Example of outstanding GPT-4 performance.



**Model Input**

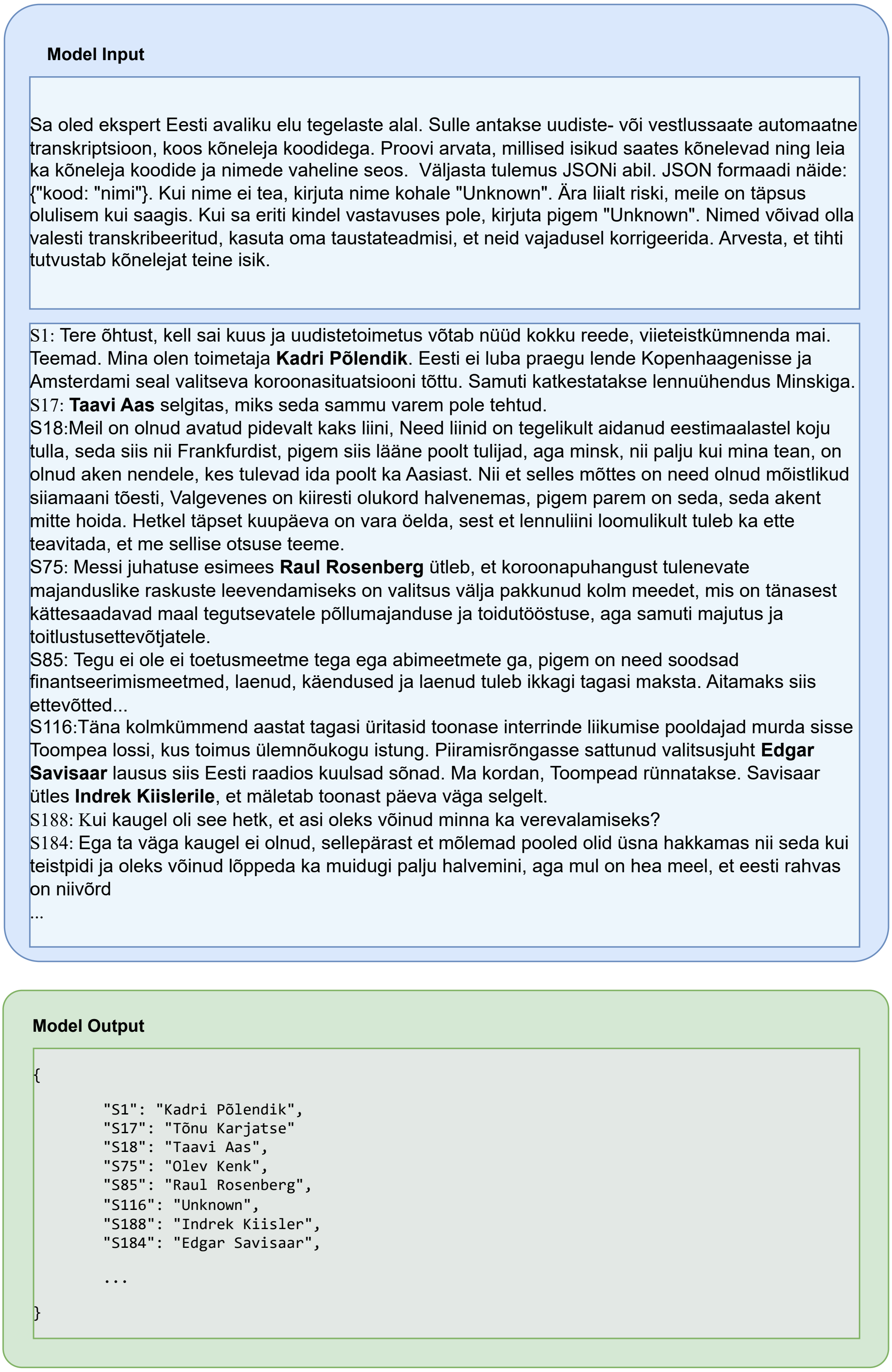
Here are some interview segment transcripts from a certain celebrity. Who do you think it might be? Please provide the top 10 guesses. Present the result as a JSON-formatted list of lists, for example: [[First Name, Second Name], [First Name1, Second Name1]]. In case you cannot identify an individual, please provide your best 10 guesses; otherwise, present an empty list.

S1: Well, you know, it's not as dramatic as you're characterizing it. You know, we've got some great PowerPC products today, and we've even got some PowerPC machines in the pipeline which we haven't introduced yet. And this is going to be a more gradual transition. I think we'll hopefully when we meet with our developers a year from today, we'll have some Intel-based Macs in the marketplace. But it's going to take maybe a two-year transition. You know, we have a good relationship with IBM, and they've got a product roadmap, and today the products are really good. But as we look out into the future, where we want to go is maybe a little bit different. We never talk about unannounced products, so I can't say. There used to be a saying at Apple....
S2: She's lovely. She's kind of, I call her my warden because she keeps me like down to the ground. She has that like invisible leash. You see them in the shopping malls and stuff, on the children and the parent, and like, if she's just like... When I was really young, I made a promise, you know, we kind of had that talk about... They bring you like housewarming gifts, and they're very polite. And she brought me these Hello Kitty chopsticks, and I still haven't opened them, and I was like 13. And I just like, I will. No, I'm not a narcissist. I'm a narcissist. That would be really over the line, having sex to your own song. That's horrible. Um, we were doing a play on Greek mythology, and I was Aphrodite, and my love interest was absent, and so I used the truth. They have been a complete support system, you know, if any if I'm ever feeling.
S3: Bubble. He saw the housing bubble. He saw that all of these mortgages are going to default, and it's weird watching the movie, too, because you think, well, who are the heroes here? Because, you know, you kind of root for them, but at the same time, it's at the expense of everyone else. Bailed out the banks. Who knows what would have happened there? I didn't really know much about it at the time. You know, you read the paper, and you watch the news, but the extent of it...
S4: I'm very instinctive. I read a script, and I'm, you know, reading it, and I've been asked to look at a particular part. So I look at it, and I just think, do I understand it? Is there something I can do? There was a very obvious way of playing this character, which would have been the sort of tough MI5 guy, but I saw something a little more interesting and more sensitive in the fact that for me, it's a character that develops a conscience, really. Comes from a very tough place, is used to being very tough, but in this particular instance, develops a conscience. I'm very, when I'm not actually shooting on the set, I don't like to hang on a set. I don't like to sit around. I mean, the reality is, an awful lot of this went on...
...

**Model Output**

```
{

        "S1": [["Steve", "Jobs"], ["Tim", "Cook"], ["Elon", "Musk"], ["Larry",
"Page"], ["Jeff", "Bezos"], ["Bill", "Gates"], ["Mark", "Zuckerberg"], ...],

        "S2": [["Miley", "Cyrus"], ["Taylor", "Swift"],  ["Selena", "Gomez"], ["Demi",
"Lovato"],  ["Ariana", "Grande"],  ["Katy", "Perry"], ...],

        "S3": [["Steve Carell"], ["Brad Pitt"], ["Christian Bale"], ["Ryan Gosling"],["Will
Ferrell"], ["Adam McKay"], ["Jonah Hill"], ...],

        "S4": [["Pierce Brosnan", "Brosnan"], ["Gary Oldman", "Oldman"], ["Ian McKellen",
"McKellen"], ["Tom Hardy", "Hardy"], ...]

...

}
```

Figure 15. Outline of LLM-based speaker identification of VoxCeleb1 dataset. The instruction prompt is slightly shortened.



**Model Input**

Sa oled ekspert Eesti avaliku elu tegelaste alal. Sulle antakse uudiste- või vestlussaate automaatne transkriptsioon, koos kõneleja koodidega. Proovi arvata, millised isikud saates kõnelevad ning leia ka kõneleja koodide ja nimede vaheline seos. Väljasta tulemus JSONi abil. JSON formaadi näide: {"kood: "nimi"}. Kui nime ei tea, kirjuta nime kohale "Unknown". Ära liialt riski, meile on täpsus olulisem kui saagis. Kui sa eriti kindel vastavuses pole, kirjuta pigem "Unknown". Nimed võivad olla valesti transkribeeritud, kasuta oma taustateadmisi, et neid vajadusel korrigeerida. Arvesta, et tihti tutvustab kõnelejat teine isik.

S1: Tere õhtust, kell sai kuus ja uudistetoimetus võtab nüüd kokku reede, viieteistkümnenda mai. Teemad. Mina olen toimetaja **Kadri Põlendik**. Eesti ei luba praegu lende Kopenhaagenisse ja Amsterdami seal valitseva koroonasituatsiooni tõttu. Samuti katkestatakse lennuühendus Minskiga.
S17: **Taavi Aas** selgitas, miks seda sammu varem pole tehtud.
S18:Meil on olnud avatud pidevalt kaks liini, Need liinid on tegelikult aidanud eestimaalastel koju tulla, seda siis nii Frankfurdist, pigem siis lääne poolt tulijad, aga minsk, nii palju kui mina tean, on olnud aken nendele, kes tulevad ida poolt ka Aasiast. Nii et selles mõttes on need olnud mõistlikud siiamaani tõesti, Valgevenes on kiiresti olukord halvenemas, pigem parem on seda, seda akent mitte hoida. Hetkel täpset kuupäeva on vara öelda, sest et lennuliini loomulikult tuleb ka ette teavitada, et me sellise otsuse teeme.
S75: Messi juhatuse esimees **Raul Rosenberg** ütleb, et koroonapuhangust tulenevate majanduslike raskuste leevendamiseks on valitsus välja pakkunud kolm meedet, mis on tänasest kättesaadavad maal tegutsevatele põllumajanduse ja toidutööstuse, aga samuti majutus ja toitlustusettevõtjatele.
S85: Tegu ei ole ei toetusmeetme tega ega abimeetmete ga, pigem on need soodsad finantseerimismeetmed, laenud, käendused ja laenud tuleb ikkagi tagasi maksta. Aitamaks siis ettevõtted...
S116:Täna kolmkümmend aastat tagasi üritasid toonase interrinde liikumise pooldajad murda sisse Toompea lossi, kus toimus ülemnõukogu istung. Piiramisrõngasse sattunud valitsusjuht **Edgar Savisaar** lausus siis Eesti raadios kuulsad sõnad. Ma kordan, Toompead rünnatakse. Savisaar ütles **Indrek Kiislerile**, et mäletab toonast päeva väga selgelt.
S188: Kui kaugel oli see hetk, et asi oleks võinud minna ka verevalamiseks?
S184: Ega ta väga kaugel ei olnud, sellepärast et mõlemad pooled olid üsna hakkamas nii seda kui teistpidi ja oleks võinud lõppeda ka muidugi palju halvemini, aga mul on hea meel, et eesti rahvas on niivõrd
...

**Model Output**

```
{

        "S1": "Kadri Põlendik",
        "S17": "Tõnu Karjatse"
        "S18": "Taavi Aas",
        "S75": "Olev Kenk",
        "S85": "Raul Rosenberg",
        "S116": "Unknown",
        "S188": "Indrek Kiisler",
        "S184": "Edgar Savisaar",

        ...

}
```

Figure 16. Outline of LLM-based speaker identification of broadcast news and multiparty conversations. The instruction prompt is slightly shortened.

# 5. Summary

Our study investigates methods to improve speaker identification accuracy by leveraging pretrained language models (LLMs). A key finding is the LLMs' capability to accurately infer speakers' full names from speech transcripts, especially when speakers are formally introduced. This practice is common in various conversational contexts such as broadcast news and discussions. Remarkably, LLMs can also identify speakers' full names even with limited introductions, likely by correlating speech content and style with their online presence. This achievement is noteworthy, particularly considering the success of Estonian—a highly inflected language with around 1 million native speakers, which may not receive primary focus in LLM development efforts. Our results suggest that this approach holds significant practical applications, including automating the annotation of diverse audio archives, benefiting academic research and media/archival management.

Our methodology for speaker identification in audio transcripts consists of three steps: speaker diarization to distinguish between speakers, speech recognition to convert audio to text, and LLM-based speaker identification to assign names. We anticipate that the advancement of multimodal generative models capable of processing both audio and text data in a unified framework will streamline these steps. This integration not only simplifies the workflow but also enhances overall transcription accuracy. These advancements have the potential to revolutionize audio transcription by offering a more efficient, accurate, and seamless method of linking speakers with their spoken words.

# References

[1] Max Mathews and Sandra Pruzansky. “Talker-recognition procedure based on analysis of variance”. In: *Journal of the Acoustical Society of America* 36 (1964), pp. 2041–2047.

[2] Tomi Kinnunen and Haizhou Li. “An overview of text-independent speaker recognition: From features to supervectors”. In: *Speech Communication* 52 (2010). URL: `https://doi.org/10.1016/j.specom.2009.08.009`.

[3] Zhongxin Bai and Xiao-Lei Zhang. “Speaker recognition based on deep learning: An overview”. In: *Neural Networks* 140 (2021), pp. 65–99. URL: `https://doi.org/10.1016/j.neunet.2021.03.004`.

[4] Đorđe Grozdić, Zoran M. Saric, and Slobodan Jovičić. “Deep neural network-based speaker embeddings for end-to-end speaker verification”. In: *Proc. International Conference on Fundamental and Applied Aspects of Speech and Language*. 2015.

[5] Musab T. S. Al-Kaltakchi et al. “Comparison of I-vector and GMM-UBM approaches to speaker identification with TIMIT and NIST 2008 databases in challenging environments”. In: *Proc. 2017 25th European Signal Processing Conference (EUSIPCO)*. 2017.

[6] Shivangi Mahto, Hitoshi Yamamoto, and Takafumi Koshinaka. “I-vector Transformation Using a Novel Discriminative Denoising Autoencoderfor Noise-robust Speaker Recognition”. In: *Proc. 2017 Interspeech*. 2017.

[7] Snyder David et al. “Deep neural network-based speaker embeddings for end-to-end speaker verification”. In: *Proc. IEEE Spoken Language Technology Workshop (SLT)*. 2016.

[8] Ondřej Novotný et al. “Analysis of DNN Speech Signal Enhancement for Robust Speaker Recognition”. In: *Computer Speech and Language* 58 (2019), pp. 403–421. URL: `https://doi.org/10.1016/j.csl.2019.06.004`.

[9] Zhifu Gao et al. “Improving Aggregation and Loss Function for Better Embedding Learning inEnd-to-End Speaker Verification System”. In: *Proc. 2019 Interspeech*. 2019.

[10] Magdalena Rybicka and Konrad Kowalczyk. “On Parameter Adaptation in Softmax-Based Cross-Entropy Loss for Improved Convergence Speed and Accuracy in DNN-Based Speaker Recognition”. In: *Proc. 2020 Interspeech*. 2020.

[11] Zhongxin Bai and Xiao-Lei Zhang. "Partial AUC Optimization Based Deep Speaker Embeddings with Class-Center Learning for Text-Independent Speaker Verification". In: *Proc. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020.

[12] Weidi Xie and Arsha Nagrani. "Utterance-level Aggregation for Speaker Recognition in the Wild". In: *Proc. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019.

[13] Xuechen Liu, Md Sahidullah, and Tomi Kinnunen. "A Comparative Re-Assessment of Feature Extractors for Deep Speaker Embeddings". In: (2020). URL: `https://doi.org/10.48550/arXiv.2007.15283`.

[14] Mirco Ravanelli and Yoshua Bengio. "Speaker Recognition from Raw Waveform with SincNet". In: *Proc. 2018 IEEE Spoken Language Technology Workshop (SLT)*. 2018.

[15] Daniel Povey, Gaofeng Cheng, and Yiming Wang. "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks". In: *Proc. Interspeech*. 2018.

[16] Ya-Qi Yu, Lei Fan, and Wu-Jun Li. "Ensemble Additive Margin Softmax for Speaker Verification". In: *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019.

[17] Weiwei Lin and Man-Wai Mak. "Wav2Spk: A Simple DNN Architecture for Learning Speaker Embeddings from Waveforms". In: *Proc. Interspeech*. 2020.

[18] Jee- Weon Jung et al. "A Complete End-to-End Speaker Verification System Using Deep Neural Networks: From Raw Signals to Verification Result". In: *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018.

[19] Jinkun Chen Weicheng Cai. "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System". In: *Proc. Odyssey*. 2018.

[20] Nanxin Chen Jesus Villalba. "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations". In: *Computer Speech and Language* 60 (2020). URL: `https://doi.org/10.1016/j.csl.2019.101026`.

[21] Ashish Vaswani et al. "Attention is all you need". In: *Proc. NeurIPS*. 2017. URL: `https://doi.org/10.48550/arXiv.1706.03762`.

[22] OpenAI. "GPT-4 Technical Report". In: *CoRR* abs/2303.08774 (2023). DOI: `10.48550/ARXIV.2303.08774`. arXiv: `2303.08774`. URL: `https://doi.org/10.48550/arXiv.2303.08774`.

[23] Andrew M. Dai and Quoc V. Le. “Semi-supervised Sequence Learning”. In: *Proc. NeurIPS*. 2015. URL: `https://doi.org/10.48550/arXiv.1511.01432`.

[24] Jacob Devlin et al. “BERT: pre-training of deep bidirectional transformers for language understanding”. In: *Proc. NAACL-HLT*. 2019. URL: `https://doi.org/10.48550/arXiv.1810.04805`.

[25] Yinhan Liu et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *CoRR* abs/1907.11692 (2019). arXiv: `1907.11692`. URL: `http://arxiv.org/abs/1907.11692`.

[26] Leonardo Canseco-Rodriguez, Lori Lamel, and Jean-Luc Gauvain. “Speaker diarization from speech transcripts”. In: *Proc. ICSLP*. Vol. 4. 2004, pp. 3–7.

[27] Julie Mauclair, Sylvain Meignier, and Yannick Estève. “Speaker diarization: about whom the speaker is talking?” In: *Proc. Speaker Odyssey 2006*. `https://hal.archives-ouvertes.fr/hal-01434121`. San Juan, Puerto Rico, 2006.

[28] Y. Estève, Sylvain Meignier, and Julie Mauclair. “Extracting true speaker identities from transcriptions”. In: *Proc. Interspeech*. Antwerp, Belgium, Aug. 2007.

[29] V. Jousse et al. “Automatic named identification of speakers using diarization and ASR systems”. In: *Proc. ICASSP*. Taipei, Taiwan, 2009, pp. 4557–4560. DOI: `10.1109/ICASSP.2009.4960644`.

[30] Elie El Khoury et al. “Combining transcription-based and acoustic-based speaker identifications for broadcast news”. In: *Proc. ICASSP*. Kyoto, Japan, 2012.

[31] Simon Petitrenaud et al. “Identification of Speakers by Name Using Belief Functions”. In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 179–188.

[32] Hervé Bredin et al. “Person instance graphs for named speaker identification in TV broadcast”. In: *Proc. Speaker Odyssey 2014*. 2014.

[33] M. Chengyuan, Patrick Nguyen, and Milind Mahajan. “Finding speaker identities with a conditional maximum entropy model”. In: *Proc. ICASSP*. Honolulu, HI, USA, Apr. 2007.

[34] P. Moschonas and C. Kotropoulos. “Multimodal Speaker Identification Based on Text and Speech”. In: *Biometrics and Identity Management*. Ed. by B. Schouten et al. Vol. 5372. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008. DOI: `10.1007/978-3-540-89991-4_11`.

[35] Efstathios Stamatatos. “Authorship verification: a review of recent advances”. In: *Research in Computing Science* 123 (2016), pp. 9–25.

[36] Steven HH Ding et al. “Learning stylometric representations for authorship analysis”. In: *IEEE Transactions on Cybernetics* 49.1 (2017), pp. 107–121.

[37] Marcelo Luiz Brocardo et al. “Authorship verification for short messages using stylometry”. In: *2013 International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE. 2013, pp. 1–6.

[38] Douglas Bagnall. “Author identification using multi-headed recurrent neural networks”. In: *arXiv preprint arXiv:1506.04891* (2015).

[39] Maël Fabien et al. “BertAA: BERT fine-tuning for authorship attribution”. In: *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*. NLP Association of India (NLPAI). Indian Institute of Technology Patna, Patna, India, 2020, pp. 127–137.

[40] Muhammad Zeeshan; Huma Qayoom; Farman Hassan. “Robust Speech Emotion Recognition System Through Novel ER-CNN and Spectral Features”. In: *Proc. ISAECT*. IEEE. 2021.

[41] Jaejin Cho et al. “Deep neural networks for emotion recognition combining audio and transcripts”. In: *arXiv preprint arXiv:1911.00432* (2019).

[42] Ayoub Ghriss et al. “Sentiment-Aware Automatic Speech Recognition Pre-Training for Enhanced Speech Emotion Recognition”. In: *Proc. ICASSP*. 2022.

[43] Bagus Tris Atmaja and Akira Sasou. “Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition”. In: *IEEE Access* 10 (2022), pp. 124396–124407. DOI: `10.1109/ACCESS.2022.3225198`.

[44] Edmilson Morais et al. “Speech emotion recognition using self-supervised features”. In: *Proc. ICASSP*. IEEE. 2022, pp. 6922–6926.

[45] Jason Wei, Yi Tay, and Rishi Bommasani. “Emergent Abilities of Large Language Models”. In: (2022).

[46] Liyizhe Peng et al. “Customising General Large Language Models for Specialised Emotion Recognition Tasks”. In: *arXiv preprint arXiv:2310.14225* (2023).

[47] Chia-Yu Hung et al. “Who Wrote it and Why? Prompting Large-Language Models for Authorship Verification”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023, pp. 14078–14084.

[48] Tae Jin Park et al. “Enhancing speaker diarization with large language models: A contextual beam search approach”. In: *arXiv preprint arXiv:2309.05248* (2023).

[49] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. “VoxCeleb: a large-scale speaker identification dataset”. In: (2017).

[50] Alec Radford et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. DOI: `10.48550/ARXIV.2212.04356`. URL: `https://arxiv.org/abs/2212.04356`.

[51] Mirco Ravanelli et al. *SpeechBrain: A General-Purpose Speech Toolkit*. arXiv:2106.04624. 2021. arXiv: `2106.04624 [eess.AS]`.

[52] Priit Käärd. "Weakly Supervised Speaker Identification System Implementation based on Estonian Public Figures". MA thesis. Tallinn University of Technolgy, 2023.

[53] Mart Karu and Tanel Alumäe. "Weakly Supervised Training of Speaker Identification Models". In: *Proc. Speaker Odyssey 2018*. 2018.

[54] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification". In: *Proc. Interspeech*. Ed. by Helen Meng, Bo Xu, and Thomas Fang Zheng. ISCA, 2020, pp. 3830–3834.

[55] Aivo Olev and Tanel Alumäe. "Estonian speech recognition and transcription editing service". In: *Baltic Journal of Modern Computing* 10.3 (2022), pp. 409–421.

[56] Arun Babu et al. "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale". In: *arXiv preprint arXiv:2111.09296* (2021).

[57] Alexis Conneau et al. "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proc. ACL*, pp. 8440–8451.

# Appendix 1 – Non-Exclusive License for Reproduction and Publication of a Graduation Thesis[1]

I Oleksandra Zamana

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis “Using Pretrained Language Models for Improved Speaker Identification”, supervised by Tanel Alumäe
    1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
    1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

08.05.2024

---

[1] The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

# Appendix 2 - Example of Full News Transcript

| Speaker | Model Input |
|---|---|
| S1 | Tere õhtust, kell sai kuus. Uudistetoimetus teeb kokkuvõtte pühapäevast, neljateistkümnendast märtsist emakeelepäevast. Stuudios on toimetaja Tõnu Karjatse. Kultuurkapital kuulutas välja kirjandusauhinna laureaadid. Lätiski kasvab rahulolematus vaktsineerimise tempoga. Peaksime kritiseerimise asemel hoopis keskenduma meditsiinitöötajate aitamisele ja viiruse leviku peatamisele, ütleb Toomas Sildam oma nädala kommentaaris. Belgia on pärast mitu kuud kestnud karme piiranguid asunud ühiskonda järk-järgult avama. Kuigi tõlkeprogrammid lähevad iga aastaga paremaks, ei suuda need lähiaastatel kindlasti asendada suulist tõlget, ütleb keeletehnoloogi. |
| S1 | Ilmateenistus lubab lund, lörtsi ja ka vihma. Õhutemperatuur on öösel miinus kaks kuni pluss üks, päeval aga kuni pluss viis kraadi ja teed on libedad. |
| S1 | Nüüd kõigest lähemalt ööpäevaga tuvastati Eestis üheksasada kolmkümmend viis koroonaviirusega nakatumist. Analüüsiti veidi üle viie tuhande proovi, neist osutus positiivseks kaheksateist koma viis protsenti. Haiglaravi vajab praegu ligi seitsesada inimest. Intensiivravi kuuskümmend seitse ööpäeva jooksul suri üheksa koroonaviirusega nakatunud inimest, neist vanim on üheksakümne kolme, noorim aga viiekümne aastane. Vaktsineeritud on koroonaviiruse vastu kaksteist protsenti täisealisest elanikkonnast. Toomas Sildami nädala kommentaar on samuti vaktsineerimise ja riigi lukkupaneku teemal. |
| S17 | Ettevõtja Indrek Kasela võttis mult sõnad, kui kirjutas sotsiaalmeedias, et lõpetame nüüd eriolukorrast jauramise ja keskendume meditsiini aitama. Oleme ise hoolsad ja ei koorma haiglaid. Saan aru, kui endine välisminister Urmas Reinsalu nõuab, et riik tuleb viia teistsugusele juhtimismudelile ja selleks on vaja kehtestada eriolukord. Reinsalu on praegu koos Isamaaga opositsioonis ja tema roll ongi valitsusele tülikas olla. Aga miks sama teeb Keskerakonna esimees Jüri Ratas praeguse peaministri Kaja Kallase koalitsioonipartner, see on raskemini mõistet. Ärme pelga eriolukorra väljakuulutamist tahab ratas otsekui kella tagasi keerata mullu kaheteistkümnenda märtsi hilisõhtusse, kui tema tollase peaministri teatas eriolukorra kehtestamisest. Nüüd on see vastuvoolu ujumine. Krista Fischer teadusnõukojast loodab, et kasu on praegustest piirangutest, mis on kahtlemata rangemad kui need, mis olid novembris ja detsembris eelmise valitsuse ajal. |
| S17 | Terviseameti peadirektor Üllar Lanno ei näe, et eriolukorra väljakuulutamine praeguse viirusekriisi lahendamisele midagi olulist juurde annaks. Enam ei ole riigi lukkupanekuks vaja Eestis eriolukorda, sest riigikogu täiendas mullu seadusi nii, et riigi saab lukku panna ka ilma eriolukorrata. Ja kui keegi otsib vastust küsimusele, kes kriisi lahendamise eest vastutab, siis see on teada. Peaminister Kaja Kallas. Ratase ja Reinsalu jutt eriolukorra vajalikkusest mõjub kahe poliitiku nostalgiana nende noorusaja järele. Ent paljudele inimestele tundub eriolukord imelise võlukepina, sest Nordstati küsitluse järgi arvab kuuskümmend üheksa protsenti vastanutest, et sarnaselt mullu kevadel tuleks praegugi kuulutada välja eriolukord. Mõistlik valitsus ei lähtu avaliku arvamuse küsitlustest, vaid teadusnõukoja ettepanekud Nende järgi. Need ministrid kehtestasidki möödunud kevadel sarnased piirangud, mis peaksid vähendama meie kõigi omavahelisi kontakte ja peatama meditsiini kriisi süvenemise. |
| S17 | Plaanilise ravi sulgumise ja haiglate kokkukukkumise eeldus on, et inimesed peavad nendest piirangutest kinni. Ettevõtja Tiit Pruuli arvates ei ole Kaja Kallase valitsuse esimene kuu olnud ehk kõige selgemate ja jõulisemate otsuste aeg. Seda ei olnud ka Jüri Ratase valitsuse viimane kuu kasvava nakatumisega. Eesti jäi lukku, kui vanemate piirangud tabasid vaid Ida-Virumaad ja Harjumaad kuid ei peatanud koroona levikut. Ent me ei pea kättemaksu tooniliselt otsima süüdlasi, aga analüüsida tuleb kindlasti, soovitab pruuli, meenutades, et süüdistamine ja analüüsimine on erinevad asjad. Mullu kevade lõpul ja suve alguses oli tegutsejate väsimus koroona esimesest lainest liiga suur, et õppetunde kaardistada ja mõtestada. Inimlikult täiesti mõistetav. Nüüd aga tuleb selleks jõud ja aeg võtta mis tähendab ka e-Eesti auditit ja mitmete IT-lahenduste kaasajastamist või hoopis ümbertegemist. Seniks aga hoiame ennast ja kaaslasi, oleme terved. |
| S1 | Läti on aprilli algusest valmis vaktsineerima kuni sada viiskümmend tuhat inimest nädalas, sellest luuakse üle kogu riigi suured vaktsineerimiskeskused. Rahulolematus senise vaktsineerimiskorralduse üle aga kasvab. Sel nädalal jõudis avalikkuseni veel kolmas juhtum, kus ametnikud on pakutud doosidest valitsusest mööda minnes keeldunud. Telefonil on nüüd Ragnar Kond, Ragnar, kuidas Läti kavatseb siis massivaktsineerimist korraldada. |

| Speaker | Model Input |
|---|---|
| S52 | Kõigepealt siis luuakse jah, valmisolek suurte keskuste näol alates esimesest aprillist, neid keskuse tuleb igale poole regioonides ja neid tuleb ka Riiga. Ja kui vaadata numbreid, siis sellest saja viiekümnest tuhandest doosist, mis siis, kui kus võimekus nädalas, Lätis saab olema umbes pool peaks siis katma just need keskused, mis luuakse lisaks veel siis perearstidele lisaks veel haiglate le, ja suurenema peab ka siis töökollektiivide ja koolide osa, kus vaktsineeritakse kohapeal ja alates järgmisest nädalast on Lätis võimalik vaktsineerida ka kodus näiteks liikumispuudega inimesi, vanemaid inimesi, kellel on võimatu minna tervishoiuasutusse, neil on võimalik nüüd helistada kas oma perearstile või, või valvetelefonile ja nad saavad siis vaktsineerida end kodus. Läti rahulolematus on seotud jah, sellega, et kui vaadata kasvõi statistikat, siis teistest Euroopa riikidest on Läti saanud neid vaktsiinidoose oluliselt vähem. Ja kuigi nüüd Läti kutsub kokku ka siis koos mõne riigiga veel Euroopa ülemkogu, et seda ebavõrdset olukorda kuidagi klaarida, siis me näeme, et väga palju probleeme on olnud ka riigi sees ja tõenäoliselt on olnud üks ametnike grupp. Praegu üritatakse mitme teenistusalase juurdluse abil selgitada, kes täpselt selle grupi siis moodustavad, kes on lihtsalt läinud valitsusest mööda ja otsustanud osad vaktsiini partiid siis tagasi lükata. Jutt on kahest Paizeri BioN vaktsiini partiist üsna suurest ja nüüd tuli siis välja ka veel Moderna. Nii et Läti üritab praegu seda niisugust mahajäämust, mis on kohalike ametnike sül tekkinud ületada ja siis vaktsineerimise tempot ikkagi tõsta ja jõuda suve lõpuks seitsmekümne protsendini täiskasvanud elanikkonnast. |
| S1 | Aitäh Ragnar. |
| S1 | Mitu kuud karmide piirangute all elanud Belgia on asunud ühiskonda taas avama. Klientide tulva alla jäänud juuksurite sõnul olid inimeste soengud lausa metsikuks kasvanud, jätkab Joosep Värk Brüsselist. |
| S75 | Eelmise aasta novembri algus oli mitmete Lääne-Euroopa riikide jaoks tõehetk, sest koroonaviiruse nakatumisnäitajad kerkisid seninägematutesse kõrgustesse. Toona suleti ühiskonnad ning veel mõne nädala taguse Eestiga võrreldes neid lahti tehtud polegi. Epidemioloogiline olukord on lubanud Belgia ühiskonda avada järk-järgult. Kuu aega tagasi said ukse tavad juuksurid, paar nädalat hiljem tatoveeringu salongid ning sellest nädalast juuksurikoolid, fotograafi, ärid ja privaatsaunad. Brüsseli kesklinna kaubanduskeskuses juuksurisalongi pidava Palma sõnul oli veebruari keskpaigas klientide huvi tohutu. |
| S75 | Meil on olnud tõesti palju kliente, kuni praeguseni töötasime tihti ilma pausideta, aga nüüd on olnud rahulikum, ütles Palma. Mõni tänav eemal juuksurina töötava Nurse sõnul olid kolme ja poole kuu pikkuse pausi jooksul klientide soengud metsikuks kasvanud. |
| S75 | Klientide juuksed on väga pikad, paljud lõikasid endale kodus ise juukseid. Töö käigus näeme väga palju moraalset, väsinud isegi psühholoogilise probleemidega inimesi ollakse palju üksi ja paljudel on keeruline sellega harjuda, sest pole sotsiaalset elu. See on raske, ütles Nurseeli tavapärasemast. Pikemaid juukseid on kohanud ka palma. |
| S75 | Jah, väga palju, eriti palju näeme seda just meeste puhul. Kui naistel on juba pikemad juuksed, siis eelistavad nad tihti neid ka pikemana hoida. Nad ütlevad, et nad on pikemate juustega harjunud, aga meestel tõepoolest on kohati väga väga-väga pikad juuksed ja kõik on pärast lõikust tulemusega rahul olnud, ütles palma. Juuksurite sissetulek on nüüd vaikselt taastumas, kuid finantsiline auk on meeletu. |
| S84 | Rahalises mõttes tabas katastroof meid kohe pärast avamist, sest pidime salongi eest ikkagi makse maksma. Erateenusepakkujatele oli see aga kingitus, mille riik justkui laenuna tegi. Nurseli sõnul on lähi. |
| S75 | Piirkonnas uksed sulgenud kahe tema tuttava juuksuri salongi, sest raha sai lihtsalt otsa. Vahepealset aega üritavad tasa teha ka tatoveerijad, kelle käed on sõna otseses mõttes tööd täis. Ülipingelisele graafikule viidates loobusid neist ajakirjanikuga rääkimast. Kõik järjekorrad olla mitmete kuude pikkused. Joosep Värk, Brüssel. |
| S1 | Tänane emakeelepäev kultuurkapital kuulutas sel puhul välja kirjanduse aastaauhindade laureaadid. Luulepreemia saab Tõnis Vilu tundekasvatus ning proosapreemia mudlumi romaan mitte ainult minu tädi Ellen. Preemia ilukirjandusliku tõlke eest võõrkeelest eesti keelde saab Triinu Tamm, kes tõlkis prantsuse keelest Patrick De Vili teose katk ja koolera ning Nansustoni teose Ingli märk eesti keelest. Võõrkeelde tõlkija. Te seast saab kirjanduse aastaauhinna Jouko Anhanen, kes tõlkis soome keelde Jaan Krossi maailma avastamise ja taevakivi. Lastekirjandus auhinna laureaat Ton Juhani püttsepp oma raamatuga on kuu kui kuldne laev. Tänasel emakeele päeval on paslik ka küsida, mis seisus on eesti keele tehnoloogiline arendamine. Indrek Kiisler uuris seda lähemalt. |

| Speaker | Model Input |
|---|---|
| S106 | Kes meist poleks olnud hädas mõne võõrkeele õppimisega, ikka küsivad lapsed, kas lõpuks pole tõesti võimalik leiutada tõlkerakendust, mis lubab rääkida meil eestlastel oma ilusas emakeeles ja meie võõrkeelsetel vestluspartneritel oleks kõrvaklappidest kosta nende emakeelne otsetõlge? Tallinna tehnikaülikooli endine rektor professor Jaak Aaviksoo on rõhutanud et sääraste nii-öelda kõrvatõlkide rakendustes peab kindlasti olemas olema ka eesti keel. See on meie tuleviku jaoks ülimalt oluline. Tallinna tehnikaülikooli keeletehnoloogia labori vanemteadur Einar Meister ütles, et eesti keel on nendes rakendustes täiesti olemas, kuid mehaaniline tõlge ei tähenda veel keelest arusaamist. |
| S112 | Masinad võivad tõlkida küll jah, vabalt või suhteliselt hästi ühes keelest teise ja noh, kas või selle Google'i masintõlke arengut vaadates siis keelte paaride arv, mida nad siis katavad, et see on järjest kasvanud ja see tõlke kvaliteet on ka läinud kogu aeg paremaks. Aga, ja on palju selliseid rakendusi, mille puhul täiesti piisab sellest, et, et see tuvastatud tekst tõlkida teise keelde. Aga me ei saa kindlasti rääkida sellest, et näiteks mingit ilukirjandusliku teksti või, või luulet võiks selliselt tõlkida. Nii et, et see oleks ka adekvaatselt |
| S112 | Mõistetav siis, et see eeldab ikkagi sellise intelligentsuse olemasolu. Ja ma ei ole väga optimist siiski selles osas, et et masinad oma |
| S112 | Intelligentsuse tasemelt jõuavad siis samale tasemele inimesega nii-öelda lähiaastatel. |
| S106 | Einar Meister rõhutas, et eesti keel on uutes keeletehnoloogia rakendustes väga hästi esindatud. |
| S112 | No näiteks meil on sellised vabavaralised rakendused, mida inglise keele puhul ka näiteks vabalt ei leia. Meie labori kodulehelt leiab sellise kõnetuvastusrakenduse, kuhu igaüks võib üles laadida oma kõnesalvestuse ja saab vastu emaili peale siis selle automaatse transkriptsiooni minu teada sellist vabavaralist rakendust inglise keele jaoks ja noh, siin ka lähikeelte jaoks soome keele puhul ma ei tea, et oleks vabavaraliselt olemas kõigile kasutada. |
| S1 | Sajusest ilmast räägib Helve Meitern. |
| S129 | Esmaspäeva öösel on pilves selgimistega ilm, sajab vihma ja lörtsi, pärast keskööd läheb sadu mitmel pool üle lumeks ja eemaldub Ida-Eestisse. On jäidet. Tuul pöördub lõunakaarest läänekaarde kaks kuni kaheksa meetrit sekundis ja õhutemperatuur on miinus üks kuni pluss üks kraad. Homme päeval madalrõhuala rüpes on meil pilves selgimistega ilm, paljudes kohtades sajab lörtsi, kohati ka lund ja saartel vihma. Päikeselisem on päev Lõuna-Eestis. Puhub läänekaare tuul kolm kuni kaheksa, puhanguti kuni kaksteist meetrit sekundis ja õhutemperatuur on pluss üks kuni pluss neli kraadi. |
| S1 | Aitäh ilmateate eest selline oligi Päevakaja emakeelepäeval mõnusat õhtut. |

| Speaker | GPT4 Prediction |
|---|---|
| S1 | Unknown |
| S17 | Toomas Sildam |
| S52 | Ragnar Kond |
| S75 | Joosep Värk |
| S84 | Unknown |
| S106 | Indrek Kiisler |
| S112 | Einar Meister |
| S129 | Helve Meitern |