Using Pretrained Language Models for Improved Speaker Identification

Oleksanda Zamana, Priit Käärd, Tanel Alumäe

Department of Software Science
Tallinn University of Technology, Estonia
olzama@taltech.ee, prkaar@taltech.ee, tanel.alumae@taltech.ee

Abstract

In this paper, we investigate improving named speaker identification with the help of pretrained language models. First, we experiment with a supervised approach where the content of each speaker's utterances in training data is used to finetune an encoder-based BERT-style language model. Next, we experiment with large generative language models, demonstrating their ability to perform zero-shot named speaker recognition using text transcripts. In both scenarios, we experiment with two languages, including VoxCeleb1 speaker identification dataset and three Estonian broadcast news and conversational datasets. We show that large language models can provide dramatic improvements to named speaker identification performance on conversational speech where speakers are introduced by their name. Furthermore, the OpenAI GPT-4 model sometimes surpasses human performance in recalling Estonian speaker names from public debate transcripts.

1. Introduction

The majority of contemporary speaker identification systems rely on the audio modality to recognize individuals. This reliance is justified given the significant advancements in accuracy these systems have experienced recently. Nonetheless, it is evident that human recognition processes are more nuanced, often incorporating both auditory cues – including spectral details, intonation, and stress – and linguistic content, such as the topics discussed and the specific words and phrase used, to identify someone familiar. For instance, a listener might subconsciously disregard an initial hypothesis about a speaker's identity based on audio cues alone if the subject matter of the conversation diverges from topics typically associated with the presumed speaker. In scenarios where the listener is unfamiliar with the person of interest, identification often hinges on the speech content, particularly when the individual's name must be deduced from introductions within the conversation. This inference can occur through direct self-introduction by the speaker or through introductions made by others participating in the interaction. Such contexts underscore the importance of linguistic content in addition to auditory cues for identifying speakers, especially in situations where prior knowledge of the individual is absent.

In this paper, we explore text-based speaker identification by leveraging various pretrained language models to assess if the verbal content from natural human interactions enhances the precision of audio-based speaker recognition. We experiment with two languages, English and Estonian, employing multiple test datasets. Initially, we use transcribed texts from speakers' utterances to construct a text-based classification framework. This is achieved by finetuning an encoder-based pretrained language model and integrating its predictions with those from the audio-based model. Subsequently, our research extends to the application of large language models (LLMs), demonstrating their capability to conduct zero-shot open-set speaker identification for well-known individuals with significant online footprints, although with restricted accuracy. The study further reveals the exceptional performance of LLMs in speaker identification tasks when provided with complete transcripts of interactions that include full-name introductions of speakers, as seen in broadcast news, talkshows and panel discussions. For example, on the test set of Estonian radio talkshows, GPT-4 improves the recall rate of speaker identification from 52% of the audio-based model to 98%, while having 100% precision.

2. Related work

The research on utilizing speech content for speaker identification has not received much attention in the past. Several studies have focused on identifying named speakers in French broadcast news. A frequently used method involves examining language patterns unique to identifying a specific speaker, capitalizing on the observation that in broadcast news, new speakers are typically introduced by the preceding speaker, introduce themselves, or are named by the following speaker. As a result, a speaker's name may be found in the previous, current, or next audio segment. Therefore, most studies have attempted to examine the words surrounding the mention of a person's name in automatic speech recognition (ASR) or manual transcripts to determine whether the name corresponds to the next, current, or previous speaker. This can be done using manually built rules [1] or machine learning methods. Semantic classification trees have been often employed for this purpose in previous works [2, 3, 4, 5]. Later studies have used deep belief functions [6] and person instance graphs [7] to combine audio and text-based speaker identification. In [8], an approach using a conditional maximum entropy model was proposed for the same task.

An alternative approach to utilize speech content for speaker identification involves examining if the speech segment's content aligns with the hypothesized speaker's known vocabulary and speaking style. In [9], probabilistic latent semantic indexing (PLSI) based topic modeling was used to estimate the topic distribution of speakers, based on their speech in training data. A closely related task to content-based speaker identification is authorship verification and attribution [10]. Its aim is to determine the authorship of a text by analyzing the stylistic characteristics and patterns occurring in it. Various methods have been proposed to capture the unique writing style of individual authors, such as n-gram frequencies, vocabulary richness, and syntactic structures [11]. Machine learning techniques, such as support vector machines (SVMs) and random forests [12], and more recently, deep neural networks [13, 14] have been employed to model the relationship between these

Table 1: 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

stylometric features and authors.

Several papers have investigated fusing audio and text modalities for the problem of emotion recognition. For example, [15] used LSTM to extract acoustic features and a convolutional model to extract information from word sequences. In recent years, many authors have also looked at using self-supervised text and audio representations for emotion recognition [16, 17], including pretrained large language models [18].

In [19] it was shown that LLMs can be used for authorship verification in a zero-shot and few-shot manner, with performance exceeding the state-of-the-art baselines. In the context of speaker recognition, LLMs have been used for speaker diarization: in [20] it was demonstrated that LLMs can predict which speakers correspond to which words in an ASR transcript, and fusing those predictions into an acoustics-only diarization system improves overall speaker attributed word error rate. We are not aware of any studies where LLMs have been used for speaker identification.

3. Supervised speaker identification using text classification

3.1. Method

In supervised speaker identification, we need training data, also known as enrollment data, for each target speaker. For audio-based speaker recognition systems, the specific content of these enrollment utterances is generally considered irrelevant, leading to their collection through standardized prompts. Conversely, in text-based speaker identification, the focus shifts to scenarios where speakers naturally converse in both the training and testing phases. It is posited that leveraging the textual content of these conversations could enhance the accuracy of audio-based speaker recognition systems.

We experiment with using pretrained BERT-like masked language models for the task of text-based speaker classification. These models are finetuned using speech transcripts of the target speakers, which may be annotated by humans or generated through automatic speech recognition (ASR) technologies.

This research explores two distinct scenarios of speaker identification: closed set and open set classification. In the closed set approach, the test set is comprised exclusively of speakers for whom training data is available. The experimentation for this scenario uses predominantly English language data from the VoxCeleb dataset. The open set scenario, which presents a more challenging and realistic environment, assumes the presence of speakers in the test set who were not included in the training dataset. For this scenario, we employ datasets from Estonian broadcast news, broadcast conversations, and recordings from a public opinion festival.

3.2. Experiments: VoxCeleb

VoxCeleb1 is a dataset collected from YouTube, comprising audio files obtained through an automated pipeline. The dataset encompasses recordings from 1251 distinct celebrities, with 61% male and 39% female speakers, predominantly from the USA or the UK.

VoxCeleb1 contains official development and test splits for speaker identification. They are constructed so that a single video recording for each speaker is selected for both development and test set, and all audio segments extracted from those videos are placed to the development/test set set.

VoxCeleb1 does not include segment transcripts. In order to perform text-based speaker recognition experiments, we transcribed the training and test data using Whisper (whisper-medium) [21]. We noticed that a small subset of VoxCeleb includes non-English speech. Therefore, in order to be able to use English-based pretrained language models, we applied Whisper in translation mode, which efficiently translates all non-English speech to English.

In speaker identification experiments, we considered the case where we have to classify the speaker based on all the segments extracted from a single recording. That is, for text-based speaker identification we pooled all segments from each recording together, resulting in 1251 test items for both development and test set. In audio-based identification, we simply averaged the posterior probability distributions for each segment of a single recordings.

We experimented with two model types for text-based speaker classification: Naive Bayes and Transformer. For both experiments, we applied data augmentation to the training data, involving sentence dropout (i.e., if the training sample contained several sentences, a random sentence was deleted). Since the amount of training data is highly unbalanced with respect to the speakers, we also applied data augmentation to ensure equal amount of training data for each speaker.

Naive Bayes model was trained on unigram word features. The initial model achieved accuracy of 2%. Data augmentation increased accuracy to 9% and stop word removal to 12%.

Next, we experimented with a Transformer model using pretrained case-sensitive version of the RoBERTa (large) [22] model as the starting point. It was finetuned for text classification in a standard way, using the augmented training data. The chosen training parameters included a learning rate of 1e-5, weight decay of 0.01, and 10 training epochs.

In assessing the performance of audio-based models, we used the SpeechBrain [23] ECAPA-TDNN model trained on Voxceleb1 and Voxceleb2 training data [1] to extract speaker embeddings for audio files. Subsequently, a logistic regression model was trained on the training split of Voxceleb1, achieving 99.8% accuracy. This corresponds to only two identification errors out of 1251 recordings.

The results of the experiments are given in Table 1. Since 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.

3.3. Experiments: Estonian broadcast and public debate speech

3.3.1. Data

For Estonian, we test our models on three datasets: radio news, radio talkshows and recordings of a public opinion festival (*Arvamusfestival*). Training data and the radio test data consist of individual program episodes scraped from the archive of Estonian Public Broadcasting (*ERR*). Most of the stored radio program episodes in the archive are manually annotated with the

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

Table 2: 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 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

Table 3: 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

names of the speakers appearing in the particular show. We used the recordings of small news clips, main evening radio news program (*Päevakaja*) and a daily conversational debate program (*Reporteritund*) for training and evaluation. These programs were selected due to their high speaker count and variability. The details of the Estonian data are given in Table 2. Most of the recordings originate from the years 2004 to 2022, but a few items date back to as far as 1959. The evening news and talkshow subsets were further split into training, development and test sets. In order to make evaluation more realistic, we first sorted each subset by date and extracted the development and test items from the end of the corresponding lists of recordings along the time axis.

To test out-of-domain performance, we also evaluate the models on the recorded sessions of the 2021 Estonian Opinion Festival (*Arvamusfestival*). It is a public festival whose mission is to enhance debate culture and civic education in the country. It consists of mostly 90-minute panels where around 5 speakers (invited based on their background) discuss some current topic. Panels include a moderator who guides the discussion and elicits audience questions and comments. This data is manually transcribed, including full names of the speakers that could be inferred by the annotator.

Since speaker annotations of the radio data are provided at a recording level and no time-aligned speaker labels are provided, we use the weakly supervised training method we proposed earlier [24] to train the audio-based speaker recognition models, with some improvements, fully described in [25]. First, instead of using i-vectors for speaker classification as in the original method, we employed the ECAPA-TDNN [26] model pretrained on VoxCeleb, with the output layer specific to Estonian data that was randomly initialized. Further, the pretrained ECAPA-TDNN backbone was finetuned with the rest of the model, using a smaller learning rate. The model covers speakers who occur at least 10 times in the dataset, which amounts to 2591 names. Note that this approach disregards the problem of different persons having the same name. However, this problem is relatively small in Estonian, with only a few cases among those 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.

All of the Estonian data was automatically diarized and transcribed using the freely available Estonian transcription system described in [27]. The speech recognition models are based on XLS-R-1B wav2vec2.0 models [28], finetuned for Estonian ASR using 761 hours of manually transcribed speech, mostly originating from broadcast conversational sources. The system has a word error rate (WER) of around 8% on broadcast conversations and even lower on broadcast news.

For training text-based speaker identification models on such weakly labelled data, the (unnamed) speakers in the training data were labelled by the weakly supervised audio-based speaker identification model (i.e., self-labelling was performed), and all transcribed speech segments corresponding to the labelled speakers were treated as reference training data for the text-based model. This also means that the name coverage of the text-based model is the same as that of the acoustic model, as the acoustic model assigns all out-of-vocabulary names to a single "unknown" speaker class.

Since the speakers in the radio test data are annotated at recording level, we also perform evaluation on that level: all models still classify individual speakers proposed by the speaker diarization process, but since we don't know the reference mapping between diarized speakers and speaker names appearing in the show, we pool all names for each individual show and compare this set to the reference set of that show, using precision and recall metrics. We tune our thresholds so that precision would be at least 95%, since incorrect speaker identification hypothesis is undesired from the application point of view.

3.3.2. Results

The text classification model for Estonian data was trained similarly to English data: we finetuned the case-sensitive version of the multilingual XLM-RoBERTa model [29] for single label sequence classification task. Training parameters remained the same: 1e-5 learning rate, 0.01 weight decay and 10 epochs. A similar data augmentation method as for VoxCeleb was used, although it did not improve the model's accuracy.

For combining the the text-based model with the audio-based model, we extracted the text embeddings for all speech turns of the textual training data. This was done by taking the output of the last hidden layer of the underlying XLM-RoBERTa model for the particular text and picking the vector corresponding to the first $[CLS]$ pseudo-token. Based on those embeddings, a speaker recognition model was trained: after normalization, the dimensionality of the embeddings were reduced to 150 and a generative classifier based on the PLDA paradigm was trained.

To improve speaker identification accuracy, especially in scenarios where the audio-based model demonstrates uncertainty, we developed a strategy leveraging both audio and text-based models. Typically, the audio-based model exhibits a high confidence level in its predictions, assigning a posterior probability greater than 0.95 to the correct speaker or favoring the special "unknown" speaker class when the actual speaker is not within the model's coverage, while lowering probabilities for all other speakers.

Our focus was to address cases where the audio-based model's certainty falls in the intermediate range. In these instances, the text-based model can offer additional insights to either confirm or refute the audio-based model's tentative speaker identification. To operationalize this approach, we processed the development and test datasets with the audio-based model,

extracting predictions where the posterior probability exceeded 1% (excluding those attributed to the "unknown" speaker category). Note that for many diarized speakers, there could be several of such predictions. We then computed the log-likehood ratio based score of the text-based LDA/PLDA model of each such speech turn and speaker pair. The posterior probabilities from the audio-based model and the scores from the text-based model were then combined, along with the binary reference labels indicating the presence or absence of the speaker in the episode, to train a logistic regression model. This model was subsequently applied to synthesize the final prediction hypotheses for the test data.

Speaker identification result on three test sets are listed in Table 2. It can be seen that the audio-based model alone reaches high precision, while the text-based model alone is lagging far behind. This outcome aligns with expectations, as the text-based model predominantly identifies speakers with high confidence only when they introduce themselves – a common practice in radio news broadcasts where reporters often mention their names. Despite the text-based model's limitations, integrating it with the audio-based model yields a modest improvement in precision on the news and opinion festival datasets, compared to using the audio-based model alone. A closer examination reveals that this enhancement primarily arises in scenarios where the audio-based model assigns a moderate posterior probability (e.g., 50%) to a speaker, and the text-based model effectively disambiguates these predictions in the correct direction. In the talkshow domain, most speakers usually speak for a much longer period in total, allowing the audio-based model to make more confident decisions, and the text-based model does not improve results.

4. Zero-shot speaker identification with LLMs

4.1. Method

In this section, we investigate speaker identification using LLMs. This is achieved by presenting the LLM with a transcript of a speaker's utterances(s) (in case of VoxCeleb experiment) or a speaker code attributed transcript of the whole broadcast/debate recording (in case of the Estonian experiment), and instructing the model to perform named speaker identification. This method can be considered unsupervised speaker identification, since no training transcripts were explicitly used for generating predictions. However, it is very likely that the model has seen texts characterizing our test speakers during pretraining, such as interview transcripts and stories about certain persons. Therefore, it is more common to call this scenario zero-shot inference, since the LLMs are not explicitly finetuned for this task, and no training examples are provided to the model in the prompt.

We used OpenAI's GPT-3.5 and GPT-4 [30] models for experiments. Although recently several freely available LLMs have been published, none possess significant capabilities for processing data in Estonian.

4.2. Experiments: VoxCeleb

In this experiment we evaluate the ability of LLMs to predict the speaker behind the transcripts of utterances, using no contextual information (such as interviewer's speech). Similarly to the supervised experiment, we only use the transcripts of the official VoxCeleb speaker identification audio segments, concatenated into one text passage per test speaker. Contrary to the supervised experiment, we do not inform the model in any way about the possible set of candidate speakers (except for one experiment).

Table 4: 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

Table 5: 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

More specifically, we preprocessed the VoxCeleb test data by concatenating individual utterances originating from an individual test video into a single string. We then employed the OpenAI API to dynamically query a LLM, utilizing the following prompt 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]

The model generated a list of guesses, which we subsequently processed and compared to the actual celebrity name to evaluate performance.

We tested OpenAI's GPT-3.5 (gpt-3.5-turbo-0613) and GPT-4 (gpt-4-0613). As illustrated in Table 4 we evaluated both Top1 accuracy and Top10 accuracy metrics. The Top1 accuracy indicates the model's success in identifying the correct celebrity as its first guess. On the other hand, Top10 accuracy measures how effectively the model could include the name of the correct celebrity within the entire list of guesses.

In the second part of these experiments we also provided GPT models with a set of top 10 celebrities that were assigned the highest posterior probabilities by the audio-based speaker identification model. The names of the top 10 celebtities were listed in a random order in the prompt, along with the appropriate instruction. This efficiently turns the task into a closed set identification task. Results are shown in Table 5. It can be seen that in 80% of the cases, GPT-4 picked the correct celebrity from the provided set.

Table 6 depicts some predictions of the GPT-4 model based on the shown input.

4.3. Experiments: Estonian broadcast and public debate speech

In the Estonian experiment, we allow the LLM to use the full transcript of the speech recording for speaker identification.

Table 6: Example predictions of the GPT-4 model with true speakers.

Input	Top 5 hypthesized 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** Taylor Lautner Peter Facinelli Kellan Lutz 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** Ariana Grande Selena Gomez Demi Lovato 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** Emma Watson Jennifer Lawrence Kristen Stewart 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 Robert Pattinson Leonardo DiCaprio Brad Pitt Tom Cruise	Ian Somerhalder

The transcript is produced by a rich transcription system that applies speaker diarization, and thus we can prepend all speaker turn transcripts with the corresponding speaker codes. Thus, the LLM can use more sources of information for speaker identification: the speaker naming patterns, as several previous works [2, 3, 4, 5, 6, 7, 8]), as well as the contents of the particular speaker turn itself.

Since the speech transcript is in Estonian, the prompt to the LLM is also in Estonian, and translates as:

> 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 whole workflow is shown in Figure 1. The transcripts of the opinion debate are relatively long, corresponding to a 90-minute conversation, and are therefore often more than 32k tokens long. When we started the experiments, the GPT-4 model had a context length of 8k tokens, and we thus processed each transcript in several parts. The prompt then includes also the speaker code $\rightarrow$ speaker name mappings predicted from the previous segment (with the appropriate instruction).

To save on OpenAI API costs, we used random 20-show subsets for broadcast news and talkshows for testing, which explains the slightly different results of the audio-based model, compared to Table 3. However, in recent months the OpenAI API costs have fallen significantly, and now processing a transcript of a 90 minute conversation costs around $0.30.

LLM-based speaker identification results are listed in Table 7. The precision and recall rates for the LLM-based systems are computed by allowing a 1-character Levenshtein distance with regard to the reference names. For the opinion festival transcripts, LLMs sometimes predict speaker names without the surname (since often the speakers are introduced without a surname). Such predictions were discarded.

As can be seen, GPT-4 (gpt-4-1106-preview) achieves remarkable precision and recall rates on all testsets. On talkshow and public debate data, it outperforms the audio-based model by a large margin. This is mostly because the audio based model has relatively low name coverage on those datasets, while the LLM infers all names directly from the transcripts. It is also evident that GPT-4 has dramatically better performance on this task than GPT-3.5 (gpt-3.5-turbo-16k-0613).

The last row in Table 7 corresponds to a combined system. This was implemented by using predictions from the audio

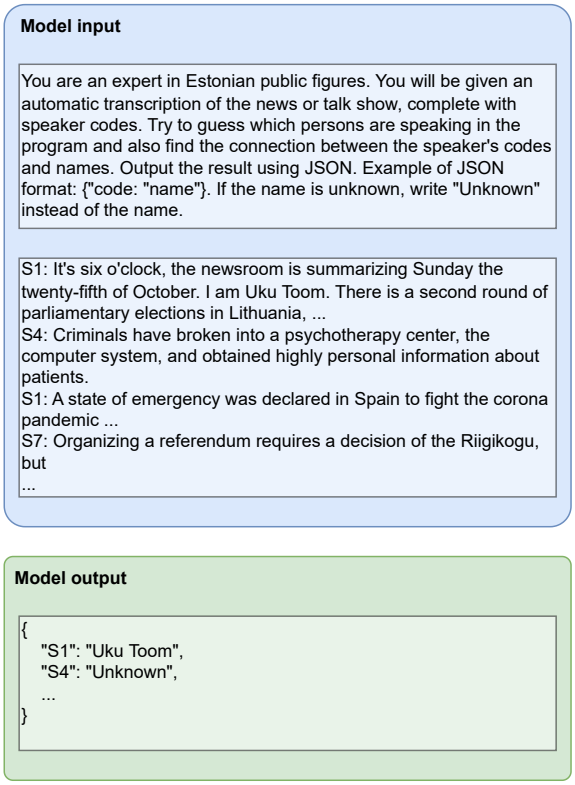

Figure 1: Outline of LLM-based speaker identification of broadcast news and multiparty conversations. Instruction prompt is slightly shortened and all interaction is translated from Estonian to English.

based model for each speaker code, if available, and using GPT-4 based name hypotheses for speakers that were not identified by the audio-based model.

As mentioned earlier, we used one-character forgiveness distance when comparing LLM-proposed speaker names with those of the reference speakers. This is needed because sometimes, even the reference transcripts contain names written with small errors in person names, especially for first names whose writing is sometimes ambiguous. The problem becomes more severe when the LLM would have to rely on ASR-generated transcripts. Table 8 shows precision and recall rates on the opinion festival recordings, when using either ASR-generated transcripts or reference transcripts, and with increasing forgiveness distance. As expected, the performance on ASR-generated transcripts is lower, but not by a large margin, especially considering that the recordings are relatively noisy and thus difficult for the ASR system.

Table 7: 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	69.5
Audio + GPT4	99.0	89.9	97.8	97.8	96.9	73.6

Table 8: 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

When analyzing GPT-4 generated speaker names for the opinion festival data, we noticed that in several cases it proposed names for speakers that were left unnamed in reference transcripts. Those cases corresponded often to scenarios when a speaker from the audience introduced her/himself by using the first name, and then continued to ask a question or make a comment about the topic of the given debate. Since the full name of the speaker was not known to the human annotator, the speaker was left unnamed. However, in some cases GPT-4 could infer the full name of the speaker based on the content of the question. It usually happens if this person is a well-known spokesperson on the given topic, or a journalist in a local newspaper where local issues are discussed. We carefully analyzed such cases, by trying to find speech samples for those particular speakers from the web and comparing them manually to the corresponding speech segments in the test data, and in most cases found the name to be correct. Therefore, it can be said that GPT-4 speaker naming abilities exceed sometimes those of a human annotator. It must be mentioned that such "hedges" by GPT-4 are not always correct and are the main cause of its less than perfect precision.

5. Discussion and conclusion

In our study, we explored methods to enhance the accuracy of speaker identification through the integration of pretrained language models. The main finding of our research is the ability of LLMs to accurately deduce speakers' full names from speech transcripts, particularly when speakers are formally introduced by name – a common practice in broadcast news, conversations, and various forms of conversational discourse. Impressively, LLMs also demonstrate a notable capacity to identify the full names of speakers, even when introductions are limited to first names, probably by relating the topic and style of their speech content with their online presence. This result is particularly notable considering its success with Estonian – a highly inflected language with around 1 million native speakers, which likely isn't a primary focus for most LLM developments. Our findings suggest that this relatively easy-to-implement approach of speaker identification has substantial practical applications, such as the potential for automating the annotation of diverse audio archives, thereby offering a valuable tool for both academic research and practical applications in media and archival management.

Our methodology for identifying speakers within audio transcripts is structured around three steps: speaker diarization to distinguish between different speakers within the audio,

speech recognition to convert audio to text, and LLM-based speaker identification to assign names to speakers. We anticipate that the evolution of multimodal generative models, which are capable of processing both audio and text data within a singular framework, will soon streamline these steps into a unified process. This integration promises not only to simplify the workflow but also to significantly enhance the overall accuracy of end-to-end rich transcriptions.

6. References

[1] Leonardo Canseco-Rodriguez, Lori Lamel, and Jean-Luc Gauvain, ," in *Proc. ICSLP*, 2004, vol. 4, pp. 3–7.

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

[3] Y. Estève, Sylvain Meignier, and Julie Mauclair, "Extracting true speaker identities from transcriptions," in *Proc. Interspeech*, Antwerp, Belgium, Aug 2007.

[4] V. Jousse, S. Petit-Renaud, S. Meignier, Y. Esteve, and C. Jacquin, "Automatic named identification of speakers using diarization and asr systems," in *Proc. ICASSP*, Taipei, Taiwan, 2009, pp. 4557–4560.

[5] Elie El Khoury, Antoine Laurent, Sylvain Meignier, and Simon Petitrenaud, "Combining transcription-based and acoustic-based speaker identifications for broadcast news," in *Proc. ICASSP*, Kyoto, Japan, 2012.

[6] Simon Petitrenaud, Vincent Jousse, Sylvain Meignier, and Yannick Estève, "Identification of speakers by name using belief functions," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods*, Berlin, Heidelberg, 2010, pp. 179–188, Springer Berlin Heidelberg.

[7] Hervé Bredin, Antoine Laurent, Achintya Sarkar, Viet-Bac Le, Sophie Rosset, and Claude Barras, "Person instance graphs for named speaker identification in TV broadcast," in *Proc. Speaker Odyssey 2014*, 2014.

[8] M. Chengyuan, Patrick Nguyen, and Milind Mahajan, "Finding speaker identities with a conditional maximum entropy model," in *Proc. ICASSP*, Honolulu, HI, USA, Apr 2007.

[9] P. Moschonas and C. Kotropoulos, *Multimodal Speaker Identification Based on Text and Speech*, vol. 5372 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2008.

[10] Efstathios Stamatatos, "Authorship verification: a review of recent advances," *Research in Computing Science*, vol. 123, pp. 9–25, 2016.

[11] Steven HH Ding, Benjamin CM Fung, Farkhund Iqbal, and William K Cheung, "Learning stylometric representations for authorship analysis," *IEEE Transactions on Cybernetics*, vol. 49, no. 1, pp. 107–121, 2017.

[12] Marcelo Luiz Brocardo, Issa Traore, Sherif Saad, and Isaac Woungang, "Authorship verification for short messages using stylometry," in *2013 International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE, 2013, pp. 1–6.

[13] Douglas Bagnall, "Author identification using multi-headed recurrent neural networks," *arXiv preprint arXiv:1506.04891*, 2015.

[14] Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida, "BertAA: BERT fine-tuning for authorship attribution," in *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, Indian Institute of Technology Patna, Patna, India, 2020, NLP Association of India (NLPAI), pp. 127–137.

[15] Jaejin Cho, Raghavendra Pappagari, Purva Kulkarni, Jesús Villalba, Yishay Carmiel, and Najim Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," *arXiv preprint arXiv:1911.00432*, 2019.

[16] Bagus Tris Atmaja and Akira Sasou, "Evaluating self-supervised speech representations for speech emotion recognition," *IEEE Access*, vol. 10, pp. 124396–124407, 2022.

[17] Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz, "Speech emotion recognition using self-supervised features," in *Proc. ICASSP*. IEEE, 2022, pp. 6922–6926.

[18] Liyizhe Peng, Zixing Zhang, Tao Pang, Jing Han, Huan Zhao, Hao Chen, and Björn W Schuller, "Customising general large language models for specialised emotion recognition tasks," *arXiv preprint arXiv:2310.14225*, 2023.

[19] Chia-Yu Hung, Zhiqiang Hu, Yujia Hu, and Roy Lee, "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.

[20] Tae Jin Park, Kunal Dhawan, Nithin Koluguri, and Jagadeesh Balam, "Enhancing speaker diarization with large language models: A contextual beam search approach," *arXiv preprint arXiv:2309.05248*, 2023.

[21] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[23] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[24] Mart Karu and Tanel Alumäe, "Weakly supervised training of speaker identification models," in *Proc. Speaker Odyssey 2018*, 2018.

[25] Priit Käärd, "Weakly supervised speaker identification system implementation based on Estonian public figures," M.S. thesis, Tallinn University of Technolgy, 2023.

[26] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, Helen Meng, Bo Xu, and Thomas Fang Zheng, Eds. 2020, pp. 3830–3834, ISCA.

[27] Aivo Olev and Tanel Alumäe, "Estonian speech recognition and transcription editing service," *Baltic Journal of Modern Computing*, vol. 10, no. 3, pp. 409–421, 2022.

[28] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[29] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. ACL*, pp. 8440–8451.

[30] OpenAI, "GPT-4 technical report," *CoRR*, vol. abs/2303.08774, 2023.