Zero-shot speaker change point detection usong large language models

Yannis Tevissen, Jérôme Boudy, Gérard Chollet, Frédéric Petitpont

ZERO-SHOT SPEAKER CHANGE POINT DETECTION USING LARGE LANGUAGE MODELS

Authors

Yannis TEVISSEN
Jérôme BOUDY
Gérard CHOLLET
Frédéric PETITPONT

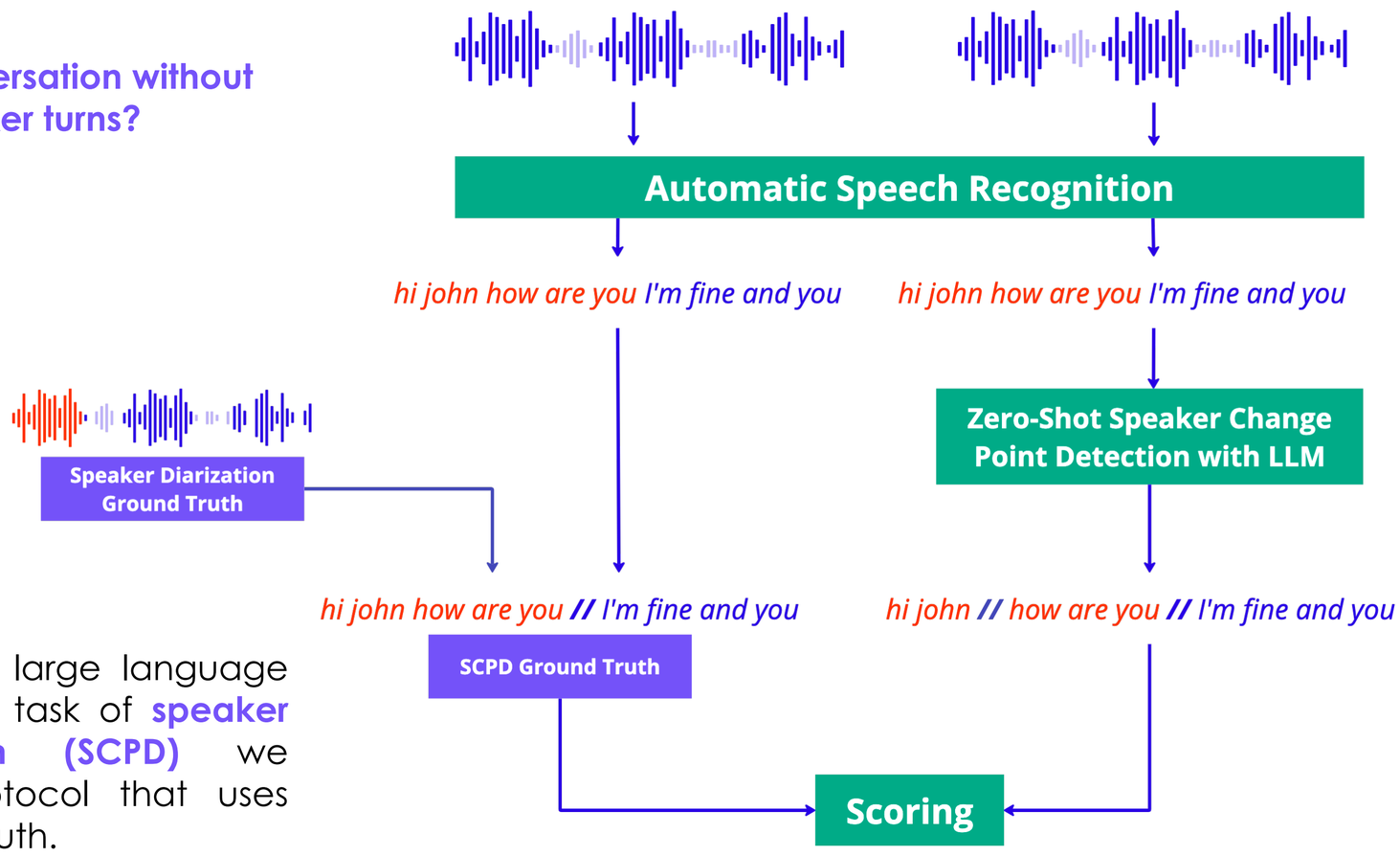
ABSTRACT

Detecting speaker changes in textual transcripts is essential for improving the readability, understanding, and analysis of multi-speaker conversations. Traditional supervised methods have shown limited success and often require a large dataset with highly accurate speaker annotations. In this study, we propose a novel **zero-shot textual speaker change point detection** approach using large language models (LLMs) to tackle the limitations of supervised techniques. Our method leverages the advanced natural language understanding capabilities of LLMs to recognize speaker change points in text **without any explicit training on speaker annotations**. We apply this technique on a diverse range of textual transcripts and study the performance of popular LLMs over it.

RESEARCH QUESTION

Can LLMs understand conversation without any indication about speaker turns?

SYSTEM PROMPT USED
You are a system designed to detect speaker change in an automatic transcript. Copy all the text I give you without correcting it and add // when you believe a speaker change happened.



To assess the capability of large language models to be good at the task of **speaker change point detection (SCPD)** we implemented a novel protocol that uses speaker diarization ground truth.

Our results show that among proprietary LLMs, only the last generation can perform this task. OpenAI GPT-4 achieves 34% accurate detection on the whole VoxConverse test set (**50 hours of recorded conversations**). Other systems were generating too many hallucinations to be scored.

	% Perfect Detection	% Close Detection	% Accurate Detection	% Missed Detection
GPT-3.5	N/A	N/A	N/A	100.0
GPT-4	11.3	22.6	34.0	66.0

Table 1. Results obtained with GPT models.

CONSISTENCY OF THE RESULTS

After passing 100 times a piece of text extracted from VoxConverse automatically generated transcription to GPT-4, we get **17 different sentence outputs**. It normally contains 2 SCPs.

1 Perfect match + 1 Missed	52% (1 alternative)
1 Close match + 1 Missed	42% (15 alternatives)
2 Missed match	6% (1 alternative)

CONCLUSION

This study shows an **emergent capability of recent large language models** to detect speaker change in a transcripted conversation.
This has to be put into perspective with the fact that we don't know all the training data of such a model.

PERSPECTIVES

AND FUTURE RESEARCH DIRECTIONS

- Speaker Change Point Detection can be used to improve **speaker diarization** (the task of determining "who spoke, when?"). Similar approaches using CNN classifiers and constrained clustering exist but none with LLMs.
- We also aim at reproducing these results with an **open-source foundation model**, possibly finetuned on some conversational data.

Key Papers

Chung, J.S., Huh, J., Nagrani, A., Afouras, T., Zisserman, A., 2020. Spot the conversation: Speaker diarisation in the wild. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2020

Anidjar, O.H., Hajaj, C., Dvir, A., Gilad, I., A Thousand Words are Worth More Than One Recording: NLP Based Speaker Change Point Detection, 2020.

Tevissen, Y., Boudy, J., Petitpont, F., 2022. The Newsbridge-Telecom SudParis VoxCeleb Speaker Recognition Challenge 2022 System Description

Contact yannis.tevissen@telecom-sudparis.eu